Unmixing Diffusion for Self-Supervised Hyperspectral Image Denoising

Haijin Zeng, Jiezhang Cao, Kai Zhang, Yongyong Chen, Hiep Luong, Wilfried Philips; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27820-27830

Hyperspectral images (HSIs) have extensive applications in various fields such as medicine agriculture and industry. Nevertheless acquiring high signal-to-noise ratio HSI poses a challenge due to narrow-band spectral filtering. Consequently the importance of HSI denoising is substantial especially for snapshot hyperspectral imaging technology. While most previous HSI denoising methods are supervised creating supervised training datasets for the diverse scenes hyperspectral cameras and scan parameters is impractical. In this work we present Diff-Unmix a self-supervised denoising method for HSI using diffusion denoising generative models. Specifically Diff-Unmix addresses the challenge of recovering noise-degraded HSI through a fusion of Spectral Unmixing and conditional abundance generation. Firstly it employs a learnable block-based spectral unmixing strategy complemented by a pure transformer-based backbone. Then we introduce a self-supervised generative diffusion network to enhance abundance maps from the spectral unmixing block. This network reconstructs noise-free Unmixing probability distributions effectively mitigating noise-induced degradations within these components. Finally the reconstructed HSI is reconstructed through unmixing reconstruction by blending the diffusion-adjusted abundance map with the spectral endmembers. Experimental results on both simulated and real-world noisy datasets show that Diff-Unmix achieves state-of-the-art performance.

****

Seeing the World through Your Eyes

Hadi Alzayer, Kevin Zhang, Brandon Feng, Christopher A. Metzler, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4864-4873

The reflective nature of the human eye is an under-appreciated source of information about what the world around us looks like. By imaging the eyes of a moving person we capture multiple views of a scene outside the camera's direct line of sight through the reflections in the eyes. In this paper we reconstruct a radiance field beyond the camera's line of sight using portrait images containing eye reflections. This task is challenging due to 1) the difficulty of accurately estimating eye poses and 2) the entangled appearance of the iris textures and the scene reflections. To address these our method jointly optimizes the cornea poses the radiance field depicting the scene and the observer's eye iris texture. We further present a regularization prior on the iris texture to improve scene reconstruction quality. Through various experiments on synthetic and real-world captures featuring people with varied eye colors and lighting conditions we demonstrate the feasibility of our approach to recover the radiance field using cornea reflections.

****

DPMesh: Exploiting Diffusion Prior for Occluded Human Mesh Recovery

Yixuan Zhu, Ao Li, Yansong Tang, Wenliang Zhao, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1101-1110

The recovery of occluded human meshes poses challenges for current methods due to the difficulty in extracting effective image features under severe occlusion. In this paper we introduce DPMesh an innovative framework for occluded human mesh recovery that capitalizes on the profound knowledge about object structure and spatial relationships embedded in a pre-trained text-to-image diffusion model. Unlike previous methods reliant on conventional backbones for vanilla feature extraction DPMesh seamlessly integrates the pre-trained denoising U-Net with potent priors as its image backbone and performs a single-step inference to provide occlusion-aware information. To enhance the perception capability for occluded poses DPMesh incorporates judicious guidance via condition injection which produces effective controls from 2D observations for the denoising U-Net. Furthermore we explore a dedicated noisy key-point reasoning approach to mitigate disturbances arising from occlusion and crowded scenarios. This strategy fully unleashes th

e perceptual capability of the diffusion prior thereby enhancing accuracy. Exten sive quantitative and qualitative experiments affirm the efficacy of our framewo rk as we outperform state-of-the-art methods on both occlusion-specific and stan dard datasets underscoring its ability to achieve precise and robust 3D human me sh recovery particularly in challenging scenarios involving occlusion and crowde d scenes. Code is available at https://github.com/EternalEvan/DPMesh.
****
Ungeneralizable Examples
Jingwen Ye, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2024, pp. 11944-11953
The training of contemporary deep learning models heavily relies on publicly ava ilable data posing a risk of unauthorized access to online data and raising conc erns about data privacy. Current approaches to creating unlearnable data involve incorporating small specially designed noises but these methods strictly limit data usability overlooking its potential usage in authorized scenarios. In this paper we extend the concept of unlearnable data to conditional data learnability and introduce UnGeneralizable Examples (UGEs). UGEs exhibit learnability for au thorized users while maintaining unlearnability for potential hackers. The prote ctor defines the authorized network and optimizes UGEs to match the gradients of the original data and its ungeneralizable version ensuring learnability. To pre vent unauthorized learning UGEs are trained by maximizing a designated distance loss in a common feature space. Additionally to further safeguard the authorized side from potential attacks we introduce additional undistillation optimization . Experimental results on multiple datasets and various networks demonstrate tha t the proposed UGEs framework preserves data usability while reducing training p erformance on hacker networks even under different types of attacks.
****
LaneCPP: Continuous 3D Lane Detection using Physical Priors
Maximilian Pittner, Joel Janai, Alexandru P. Condurache; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 106 39-10648
Monocular 3D lane detection has become a fundamental problem in the context of a utonomous driving which comprises the tasks of finding the road surface and loca ting lane markings. One major challenge lies in a flexible but robust line repre sentation capable of modeling complex lane structures while still avoiding unpre dictable behavior. While previous methods rely on fully data-driven approaches w e instead introduce a novel approach LaneCPP that uses a continuous 3D lane dete ction model leveraging physical prior knowledge about the lane structure and roa d geometry. While our sophisticated lane model is capable of modeling complex ro ad structures it also shows robust behavior since physical constraints are incor porated by means of a regularization scheme that can be analytically applied to our parametric representation. Moreover we incorporate prior knowledge about the road geometry into the 3D feature space by modeling geometry-aware spatial feat ures guiding the network to learn an internal road surface representation. In ou r experiments we show the benefits of our contributions and prove the meaningful ness of using priors to make 3D lane detection more robust. The results show tha t LaneCPP achieves state-of-the-art performance in terms of F-Score and geometri c errors.
****
CityDreamer: Compositional Generative Model of Unbounded 3D Cities
Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, Ziwei Liu; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9666-9675
3D city generation is a desirable yet challenging task since humans are more sen sitive to structural distortions in urban environments. Additionally generating 3D cities is more complex than 3D natural scenes since buildings as objects of t he same class exhibit a wider range of appearances compared to the relatively co nsistent appearance of objects like trees in natural scenes. To address these ch allenges we propose CityDreamer a compositional generative model designed specif ically for unbounded 3D cities. Our key insight is that 3D city generation shoul d be a composition of different types of neural fields: 1) various building inst

ances and 2) background stuff such as roads and green lands. Specifically we ado
pt the bird's eye view scene representation and employ a volumetric render for b
oth instance-oriented and stuff-oriented neural fields. The generative hash grid
 and periodic positional embedding are tailored as scene parameterization to sui
t the distinct characteristics of building instances and background stuff. Furth
ermore we contribute a suite of CityGen Datasets including OSM and GoogleEarth w
hich comprises a vast amount of real-world city imagery to enhance the realism o
f the generated 3D cities both in their layouts and appearances. CityDreamer ach
ieves state-of-the-art performance not only in generating realistic 3D cities bu
t also in localized editing within the generated cities.
****
HEAL-SWIN: A Vision Transformer On The Sphere
Oscar Carlsson, Jan E. Gerken, Hampus Linander, Heiner Spieß, Fredrik Ohlsson, C
hristoffer Petersson, Daniel Persson; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6067-6077
High-resolution wide-angle fisheye images are becoming more and more important f
or robotics applications such as autonomous driving. However using ordinary conv
olutional neural networks or vision transformers on this data is problematic due
 to projection and distortion losses introduced when projecting to a rectangular
 grid on the plane. We introduce the HEAL-SWIN transformer which combines the hi
ghly uniform Hierarchical Equal Area iso-Latitude Pixelation (HEALPix) grid used
 in astrophysics and cosmology with the Hierarchical Shifted-Window (SWIN) trans
former to yield an efficient and flexible model capable of training on high-reso
lution distortion-free spherical data. In HEAL-SWIN the nested structure of the
HEALPix grid is used to perform the patching and windowing operations of the SWI
N transformer enabling the network to process spherical representations with min
imal computational overhead. We demonstrate the superior performance of our mode
l on both synthetic and real automotive datasets as well as a selection of other
 image datasets for semantic segmentation depth regression and classification ta
sks. Our code is publicly available.
****
3D Paintbrush: Local Stylization of 3D Shapes with Cascaded Score Distillation
Dale Decatur, Itai Lang, Kfir Aberman, Rana Hanocka; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4473-44
83
We present 3D Paintbrush a technique for automatically texturing local semantic
regions on meshes via text descriptions. Our method is designed to operate direc
tly on meshes producing texture maps which seamlessly integrate into standard gr
aphics pipelines. We opt to simultaneously produce a localization map (to specif
y the edit region) and a texture map which conforms to it. This approach improve
s the quality of both the localization and the stylization. To enhance the detai
ls and resolution of the textured area we leverage multiple stages of a cascaded
 diffusion model to supervise our local editing technique with generative priors
 learned from images at different resolutions. Our technique referred to as Casc
aded Score Distillation (CSD) simultaneously distills scores at multiple resolut
ions in a cascaded fashion enabling control over both the granularity and global
 understanding of the supervision. We demonstrate the effectiveness of 3D Paintb
rush to locally texture different semantic regions on a variety of shapes.
****
Test-Time Linear Out-of-Distribution Detection
Ke Fan, Tong Liu, Xingyu Qiu, Yikai Wang, Lian Huai, Zeyu Shangguan, Shuang Gou,
 Fengjian Liu, Yuqian Fu, Yanwei Fu, Xingqun Jiang; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23752-23
761
Out-of-Distribution (OOD) detection aims to address the excessive confidence pre
diction by neural networks by triggering an alert when the input sample deviates
 significantly from the training distribution (in-distribution) indicating that
the output may not be reliable. Current OOD detection approaches explore all kin
ds of cues to identify OOD data such as finding irregular patterns in the featur
e space logit space gradient space or the raw image space. Surprisingly we obser

ve a linear trend between the OOD score produced by current OOD detection algorithms and the network features on several datasets. We conduct a thorough investigation theoretically and empirically to analyze and understand the meaning of such a linear trend in OOD detection. This paper proposes a Robust Test-time Linear method (RTL) to utilize such linear trends like a `free lunch' when we have a batch of data to perform OOD detection. By using a simple linear regression as a test time adaptation we can make a more precise OOD prediction. We further propose an online variant of the proposed method which achieves promising performance and is more practical for real applications. Theoretical analysis is given to prove the effectiveness of our methods. Extensive experiments on several OOD datasets show the efficacy of RTL for OOD detection tasks significantly improving the results of base OOD detectors. Project will be available at https://github.com/kfan21/RTL.
****

Guided Slot Attention for Unsupervised Video Object Segmentation
Minhyeok Lee, Suhwan Cho, Dogyoon Lee, Chaewon Park, Jungho Lee, Sangyoun Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3807-3816
Unsupervised video object segmentation aims to segment the most prominent object in a video sequence. However the existence of complex backgrounds and multiple foreground objects make this task challenging. To address this issue we propose a guided slot attention network to reinforce spatial structural information and obtain better foreground-background separation. The foreground and background slots which are initialized with query guidance are iteratively refined based on interactions with template information. Furthermore to improve slot-template interaction and effectively fuse global and local features in the target and reference frames K-nearest neighbors filtering and a feature aggregation transformer are introduced. The proposed model achieves state-of-the-art performance on two popular datasets. Additionally we demonstrate the robustness of the proposed model in challenging scenes through various comparative experiments.
****

Unsupervised Blind Image Deblurring Based on Self-Enhancement
Lufei Chen, Xiangpeng Tian, Shuhua Xiong, Yinjie Lei, Chao Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25691-25700
Significant progress in image deblurring has been achieved by deep learning methods especially the remarkable performance of supervised models on paired synthetic data. However real-world quality degradation is more complex than synthetic datasets and acquiring paired data in real-world scenarios poses significant challenges. To address these challenges we propose a novel unsupervised image deblurring framework based on self-enhancement. The framework progressively generates improved pseudo-sharp and blurry image pairs without the need for real paired datasets and the generated image pairs with higher qualities can be used to enhance the performance of the reconstructor. To ensure the generated blurry images are closer to the real blurry images we propose a novel re-degradation principal component consistency loss which enforces the principal components of the generated low-quality images to be similar to those of re-degraded images from the original sharp ones. Furthermore we introduce the self-enhancement strategy that significantly improves deblurring performance without increasing the computational complexity of network during inference. Through extensive experiments on multiple real-world blurry datasets we demonstrate the superiority of our approach over other state-of-the-art unsupervised methods.
****

Action Detection via an Image Diffusion Process
Lin Geng Foo, Tianjiao Li, Hossein Rahmani, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18351-18361
Action detection aims to localize the starting and ending points of action instances in untrimmed videos and predict the classes of those instances. In this paper we make the observation that the outputs of the action detection task can be

formulated as images. Thus from a novel perspective we tackle action detection via a three-image generation process to generate starting point ending point and action-class predictions as images via our proposed Action Detection Image Diffusion (ADI-Diff) framework. Furthermore since our images differ from natural images and exhibit special properties we further explore a Discrete Action-Detection Diffusion Process and a Row-Column Transformer design to better handle their processing. Our ADI-Diff framework achieves state-of-the-art results on two widely-used datasets.
****
Programmable Motion Generation for Open-Set Motion Control Tasks
Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1399-1408
Character animation in real-world scenarios necessitates a variety of constraints such as trajectories key-frames interactions etc. Existing methodologies typically treat single or a finite set of these constraint(s) as separate control tasks. These methods are often specialized and the tasks they address are rarely extendable or customizable. We categorize these as solutions to the close-set motion control problem. In response to the complexity of practical motion control we propose and attempt to solve the open-set motion control problem. This problem is characterized by an open and fully customizable set of motion control tasks. To address this we introduce a new paradigm programmable motion generation. In this paradigm any given motion control task is broken down into a combination of atomic constraints. These constraints are then programmed into an error function that quantifies the degree to which a motion sequence adheres to them. We utilize a pre-trained motion generation model and optimize its latent code to minimize the error function of the generated motion. Consequently the generated motion not only inherits the prior of the generative model but also satisfies the requirements of the compounded constraints. Our experiments demonstrate that our approach can generate high-quality motions when addressing a wide range of unseen tasks. These tasks encompass motion control by motion dynamics geometric constraints physical laws interactions with scenes objects or the character's own body parts etc. All of these are achieved in a unified approach without the need for ad-hoc paired training data collection or specialized network designs. During the programming of novel tasks we observed the emergence of new skills beyond those of the prior model. With the assistance of large language models we also achieved automatic programming. We hope that this work will pave the way for the motion control of general AI agents.
****
SCE-MAE: Selective Correspondence Enhancement with Masked Autoencoder for Self-Supervised Landmark Estimation
Kejia Yin, Varshanth Rao, Ruowei Jiang, Xudong Liu, Parham Aarabi, David B. Lindell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1313-1322
Self-supervised landmark estimation is a challenging task that demands the formation of locally distinct feature representations to identify sparse facial landmarks in the absence of annotated data. To tackle this task existing state-of-the-art (SOTA) methods (1) extract coarse features from backbones that are trained with instance-level self-supervised learning (SSL) paradigms which neglect the dense prediction nature of the task (2) aggregate them into memory-intensive hypercolumn formations and (3) supervise lightweight projector networks to naively establish full local correspondences among all pairs of spatial features. In this paper we introduce SCE-MAE a framework that (1) leverages the MAE [??] a region-level SSL method that naturally better suits the landmark prediction task (2) operates on the vanilla feature map instead of on expensive hypercolumns and (3) employs a Correspondence Approximation and Refinement Block (CARB) that utilizes a simple density peak clustering algorithm and our proposed Locality-Constrained Repellence Loss to directly hone only select local correspondences. We demonstrate through extensive experiments that SCE-MAE is highly effective and robust outperforming existing SOTA methods by large margins of 20%-44% on the landmark m

atching and 9%-15% on the landmark detection tasks.
****
LAKE-RED: Camouflaged Images Generation by Latent Background Knowledge Retrieval-Augmented Diffusion
Pancheng Zhao, Peng Xu, Pengda Qin, Deng-Ping Fan, Zhicheng Zhang, Guoli Jia, Bowen Zhou, Jufeng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4092-4101
Camouflaged vision perception is an important vision task with numerous practical applications. Due to the expensive collection and labeling costs this community struggles with a major bottleneck that the species category of its datasets is limited to a small number of object species. However the existing camouflaged generation methods require specifying the background manually thus failing to extend the camouflaged sample diversity in a low-cost manner. In this paper we propose a Latent Background Knowledge Retrieval-Augmented Diffusion (LAKE-RED) for camouflaged image generation. To our knowledge our contributions mainly include: (1) For the first time we propose a camouflaged generation paradigm that does not need to receive any background inputs. (2) Our LAKE-RED is the first knowledge retrieval-augmented method with interpretability for camouflaged generation in which we propose an idea that knowledge retrieval and reasoning enhancement are separated explicitly to alleviate the task-specific challenges. Moreover our method is not restricted to specific foreground targets or backgrounds offering a potential for extending camouflaged vision perception to more diverse domains. (3) Experimental results demonstrate that our method outperforms the existing approaches generating more realistic camouflage images.
****
TIGER: Time-Varying Denoising Model for 3D Point Cloud Generation with Diffusion Process
Zhiyuan Ren, Minchul Kim, Feng Liu, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9462-9471
Recently diffusion models have emerged as a new powerful generative method for 3D point cloud generation tasks. However few works study the effect of the architecture of the diffusion model in the 3D point cloud resorting to the typical UNet model developed for 2D images. Inspired by the wide adoption of Transformers we study the complementary role of convolution (from UNet) and attention (from Transformers). We discover that their respective importance change according to the timestep in the diffusion process. At early stage attention has an outsized influence because Transformers are found to generate the overall shape more quickly and at later stages when adding fine detail convolution starts having a larger impact on the generated point cloud's local surface quality. In light of this observation we propose a time-varying two-stream denoising model combined with convolution layers and transformer blocks. We generate an optimizable mask from each timestep to reweigh global and local features obtaining time-varying fused features. Experimentally we demonstrate that our proposed method quantitatively outperforms other state-of-the-art methods regarding visual quality and diversity. Code is avaiable github.com/Zhiyuan-R/Tiger-Time-varying-Diffusion-Model-for-Point-Cloud-Generation.
****
ConTex-Human: Free-View Rendering of Human from a Single Image with Texture-Consistent Synthesis
Xiangjun Gao, Xiaoyu Li, Chaopeng Zhang, Qi Zhang, Yanpei Cao, Ying Shan, Long Quan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10084-10094
In this work we propose a method to address the challenge of rendering a 3D human from a single image in a free-view manner. Some existing approaches could achieve this by using generalizable pixel-aligned implicit fields to reconstruct a textured mesh of a human or by employing a 2D diffusion model as guidance with the Score Distillation Sampling (SDS) method to lift the 2D image into 3D space. However a generalizable implicit field often results in an over-smooth texture field while the SDS method tends to lead to a texture-inconsistent novel view with the input image. In this paper we introduce a texture-consistent back view synt

hesis method that could transfer the reference image content to the back view th
rough depth-guided mutual self-attention. With this method we could achieve high
-fidelity and texture-consistent human rendering from a single image. Moreover t
o alleviate the color distortion that occurs in the side region we propose a vi
sibility-aware patch consistency regularization combined with the synthesized b
ack view texture. Experiments conducted on both real and synthetic data demonstr
ate the effectiveness of our method and show that our approach outperforms previ
ous baseline methods.
****
UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity
Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang
, Changxin Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2024, pp. 22010-22019
Existing text-based person retrieval datasets often have relatively coarse-grain
ed text annotations. This hinders the model to comprehend the fine-grained seman
tics of query texts in real scenarios. To address this problem we contribute a n
ew benchmark named UFineBench for text-based person retrieval with ultra-fine gr
anularity. Firstly we construct a new dataset named UFine6926. We collect a larg
e number of person images and manually annotate each image with two detailed tex
tual descriptions averaging 80.8 words each. The average word count is three to
four times that of the previous datasets. In addition of standard in-domain eval
uation we also propose a special evaluation paradigm more representative of real
 scenarios. It contains a new evaluation set with cross domains cross textual gr
anularity and cross textual styles named UFine3C and a new evaluation metric for
 accurately measuring retrieval ability named mean Similarity Distribution (mSD)
. Moreover we propose CFAM a more efficient algorithm especially designed for te
xt-based person retrieval with ultra fine-grained texts. It achieves fine granul
arity mining by adopting a shared cross-modal granularity decoder and hard negat
ive match mechanism. With standard in-domain evaluation CFAM establishes competi
tive performance across various datasets especially on our ultra fine-grained UF
ine6926. Furthermore by evaluating on UFine3C we demonstrate that training on ou
r UFine6926 significantly improves generalization to real scenarios compared wit
h other coarse-grained datasets. The dataset and code will be made publicly avai
lable at https://github.com/Zplusdragon/UFineBench.
****
Efficient Hyperparameter Optimization with Adaptive Fidelity Identification
Jiantong Jiang, Zeyi Wen, Atif Mansoor, Ajmal Mian; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26181-26
190
Hyperparameter Optimization and Neural Architecture Search are powerful in attai
ning state-of-the-art machine learning models with Bayesian Optimization (BO) st
anding out as a mainstream method. Extending BO into the multi-fidelity setting
has been an emerging research topic in this field but faces the challenge of det
ermining an appropriate fidelity for each hyperparameter configuration to fit th
e surrogate model. To tackle the challenge we propose a multi-fidelity BO method
 named FastBO which excels in adaptively deciding the fidelity for each configur
ation and providing strong performance while ensuring efficient resource usage.
These advantages are achieved through our proposed techniques based on the conce
pts of efficient point and saturation point for each configuration which can be
obtained from the empirical learning curve of the configuration estimated from e
arly observations. Extensive experiments demonstrate FastBO's superior anytime p
erformance and efficiency in identifying high-quality configurations and archite
ctures. We also show that our method provides a way to extend any single-fidelit
y method to the multi-fidelity setting highlighting the wide applicability of ou
r approach.
****
ASH: Animatable Gaussian Splats for Efficient and Photoreal Human Rendering
Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, Marc Habermann; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 1165-1175

Real-time rendering of photorealistic and controllable human avatars stands as a cornerstone in Computer Vision and Graphics. While recent advances in neural implicit rendering have unlocked unprecedented photorealism for digital avatars real-time performance has mostly been demonstrated for static scenes only. To address this we propose ASH an animatable Gaussian splatting approach for photorealistic rendering of dynamic humans in real time. We parameterize the clothed human as animatable 3D Gaussians which can be efficiently splatted into image space to generate the final rendering. However naively learning the Gaussian parameters in 3D space poses a severe challenge in terms of compute. Instead we attach the Gaussians onto a deformable character model and learn their parameters in 2D texture space which allows leveraging efficient 2D convolutional architectures that easily scale with the required number of Gaussians. We benchmark ASH with competing methods on pose-controllable avatars demonstrating that our method outperforms existing real-time methods by a large margin and shows comparable or even better results than offline methods.
****
Focus on Hiders: Exploring Hidden Threats for Enhancing Adversarial Training
Qian Li, Yuxiao Hu, Yinpeng Dong, Dongxiao Zhang, Yuntian Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24442-24451
Adversarial training is often formulated as a min-max problem however concentrating only on the worst adversarial examples causes alternating repetitive confusion of the model i.e. previously defended or correctly classified samples are not defensible or accurately classifiable in subsequent adversarial training. We characterize such non-ignorable samples as "hiders" which reveal the hidden high-risk regions within the secure area obtained through adversarial training and prevent the model from finding the real worst cases. We demand the model to prevent hiders when defending against adversarial examples for improving accuracy and robustness simultaneously. By rethinking and redefining the min-max optimization problem for adversarial training we propose a generalized adversarial training algorithm called Hider-Focused Adversarial Training (HFAT). HFAT introduces the iterative evolution optimization strategy to simplify the optimization problem and employs an auxiliary model to reveal hiders effectively combining the optimization directions of standard adversarial training and prevention hiders. Furthermore we introduce an adaptive weighting mechanism that facilitates the model in adaptively adjusting its focus between adversarial examples and hiders during different training periods. We demonstrate the effectiveness of our method based on extensive experiments and ensure that HFAT can provide higher robustness and accuracy. We will release the source code upon publication.
****
ArtAdapter: Text-to-Image Style Transfer using Multi-Level Style Encoder and Explicit Adaptation
Dar-Yen Chen, Hamish Tennent, Ching-Wen Hsu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8619-8628
This work introduces ArtAdapter a transformative text-to-image (T2I) style transfer framework that transcends traditional limitations of color brushstrokes and object shape capturing high-level style elements such as composition and distinctive artistic expression. The integration of a multi-level style encoder with our proposed explicit adaptation mechanism enables ArtAdapter to achieve unprecedented fidelity in style transfer ensuring close alignment with textual descriptions. Additionally the incorporation of an Auxiliary Content Adapter (ACA) effectively separates content from style alleviating the borrowing of content from style references. Moreover our novel fast finetuning approach could further enhance zero-shot style representation while mitigating the risk of overfitting. Comprehensive evaluations confirm that ArtAdapter surpasses current state-of-the-art methods.
****
GoodSAM: Bridging Domain and Capacity Gaps via Segment Anything Model for Distortion-aware Panoramic Semantic Segmentation
Weiming Zhang, Yexin Liu, Xu Zheng, Lin Wang; Proceedings of the IEEE/CVF Confer

ence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28264-28273
This paper tackles a novel yet challenging problem: how to transfer knowledge from the emerging Segment Anything Model (SAM) -- which reveals impressive zero-shot instance segmentation capacity -- to learn a compact panoramic semantic segmentation model i.e. student without requiring any labeled data. This poses considerable challenges due to SAM's inability to provide semantic labels and the large capacity gap between SAM and the student. To this end we propose a novel framework called GoodSAM that introduces a teacher assistant (TA) to provide semantic information integrated with SAM to generate ensemble logits to achieve knowledge transfer. Specifically we propose a Distortion-Aware Rectification (DAR) module that first addresses the distortion problem of panoramic images by imposing prediction-level consistency and boundary enhancement. This subtly enhances TA's prediction capacity on panoramic images. DAR then incorporates a cross-task complementary fusion block to adaptively merge the predictions of SAM and TA to obtain more reliable ensemble logits. Moreover we introduce a Multi-level Knowledge Adaptation (MKA) module to efficiently transfer the multi-level feature knowledge from TA and ensemble logits to learn a compact student model. Extensive experiments on two benchmarks show that our GoodSAM achieves a remarkable +3.75% mIoU improvement over the state-of-the-art (SOTA) domain adaptation methods e.g. [41]. Also our most lightweight model achieves comparable performance to the SOTA methods with only 3.7M parameters.
****
DYSON: Dynamic Feature Space Self-Organization for Online Task-Free Class Incremental Learning
Yuhang He, Yingjie Chen, Yuhan Jin, Songlin Dong, Xing Wei, Yihong Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23741-23751
In this paper we focus on a challenging Online Task-Free Class Incremental Learning (OTFCIL) problem. Different from the existing methods that continuously learn the feature space from data streams we propose a novel compute-and-align paradigm for the OTFCIL. It first computes an optimal geometry i.e. the class prototype distribution for classifying existing classes and updates it when new classes emerge and then trains a DNN model by aligning its feature space to the optimal geometry. To this end we develop a novel Dynamic Neural Collapse (DNC) algorithm to compute and update the optimal geometry. The DNC expands the geometry when new classes emerge without loss of the geometry optimality and guarantees the drift distance of old class prototypes with an explicit upper bound. Then we propose a novel Dynamic feature space Self-Organization (DYSON) method containing three major components including 1) a feature extractor 2) a Dynamic Feature-Geometry Alignment (DFGA) module aligning the feature space to the optimal geometry computed by DNC and 3) a training-free class-incremental classifier derived from the DNC geometry. Experimental comparison results on four benchmark datasets including CIFAR10 CIFAR100 CUB200 and CoRe50 demonstrate the efficiency and superiority of the DYSON method. The source code is provided in the supplementary material.
****
Streaming Dense Video Captioning
Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18243-18252
An ideal model for dense video captioning -- predicting captions localized temporally in a video -- should be able to handle long input videos predict rich detailed textual descriptions and be able to produce outputs before processing the entire video. Current state-of-the-art models however process a fixed number of downsampled frames and make a single full prediction after seeing the whole video. We propose a streaming dense video captioning model that consists of two novel components: First we propose a new memory module based on clustering incoming tokens which can handle arbitrarily long videos as the memory is of a fixed size. Second we develop a streaming decoding algorithm that enables our model to make predictions before the entire video has been processed. Our model achieves this

streaming ability and significantly improves the state-of-the-art on three dense video captioning benchmarks: ActivityNet YouCook2 and ViTT. Our code is released at https://github.com/google-research/scenic.
****
Rethinking Inductive Biases for Surface Normal Estimation
Gwangbin Bae, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9535-9545
Despite the growing demand for accurate surface normal estimation models existing methods use general-purpose dense prediction models adopting the same inductive biases as other tasks. In this paper we discuss the inductive biases needed for surface normal estimation and propose to (1) utilize the per-pixel ray direction and (2) encode the relationship between neighboring surface normals by learning their relative rotation. The proposed method can generate crisp - yet piecewise smooth - predictions for challenging in-the-wild images of arbitrary resolution and aspect ratio. Compared to a recent ViT-based state-of-the-art model our method shows a stronger generalization ability despite being trained on an orders of magnitude smaller dataset. The code is available at https://github.com/baegwangbin/DSINE.
****
Event-based Structure-from-Orbit
Ethan Elms, Yasir Latif, Tae Ha Park, Tat-Jun Chin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19541-19550
Event sensors offer high temporal resolution visual sensing which makes them ideal for perceiving fast visual phenomena without suffering from motion blur. Certain applications in robotics and vision-based navigation require 3D perception of an object undergoing circular or spinning motion in front of a static camera such as recovering the angular velocity and shape of the object. The setting is equivalent to observing a static object with an orbiting camera. In this paper we propose event-based structure-from-orbit (eSfO) where the aim is to simultaneously reconstruct the 3D structure of a fast spinning object observed from a static event camera and recover the equivalent orbital motion of the camera. Our contributions are threefold: since state-of-the-art event feature trackers cannot handle periodic self-occlusion due to the spinning motion we develop a novel event feature tracker based on spatio-temporal clustering and data association that can better track the helical trajectories of valid features in the event data. The feature tracks are then fed to our novel factor graph-based structure-from-orbit back-end that calculates the orbital motion parameters (e.g. spin rate relative rotational axis) that minimize the reprojection error. For evaluation we produce a new event dataset of objects under spinning motion. Comparisons against ground truth indicate the efficacy of eSfO.
****
LED: A Large-scale Real-world Paired Dataset for Event Camera Denoising
Yuxing Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25637-25647
Event camera has significant advantages in capturingdynamic scene information while being prone to noise interferenceparticularly in challenging conditions like lowthreshold and low illumination. However most existing researchfocuses on gentle situations hindering event cameraapplications in realistic complex scenarios. To tackle thislimitation and advance the field we construct a new pairedreal-world event denoising dataset (LED) including 3K sequenceswith 18K seconds of high-resolution (1200*680)event streams and showing three notable distinctions comparedto others: diverse noise levels and scenes largerscalewith high-resolution and high-quality GT. Specificallyit contains stepped parameters and varying illuminationwith diverse scenarios. Moreover based on theproperty of noise events inconsistency and signal eventsconsistency we propose a novel effective denoising framework(DED) using homogeneous dual events to generate theGT with better separating noise from the raw. Furthermorewe design a bio-inspired baseline leveraging Leaky-Integrate-and-Fire (LIF) neurons with dynamic thresholdsto realize accurate denoising. The experimental resultsdemonstrate that the remarkable performanc

e of the proposedapproach on different datasets.The dataset and codeare at https ://github.com/Yee-Sing/led.
****
Fair Federated Learning under Domain Skew with Local Consistency and Domain Dive rsity
Yuhang Chen, Wenke Huang, Mang Ye; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2024, pp. 12077-12086
Federated learning (FL) has emerged as a new paradigm for privacy-preserving col laborative training. Under domain skew the current FL approaches are biased and face two fairness problems. 1) Parameter Update Conflict: data disparity among c lients leads to varying parameter importance and inconsistent update directions. These two disparities cause important parameters to potentially be overwhelmed by unimportant ones of dominant updates. It consequently results in significant performance decreases for lower-performing clients. 2) Model Aggregation Bias: e xisting FL approaches introduce unfair weight allocation and neglect domain dive rsity. It leads to biased model convergence objective and distinct performance a mong domains. We discover a pronounced directional update consistency in Federat ed Learning and propose a novel framework to tackle above issues. First leveragi ng the discovered characteristic we selectively discard unimportant parameter up dates to prevent updates from clients with lower performance overwhelmed by unim portant parameters resulting in fairer generalization performance. Second we pro pose a fair aggregation objective to prevent global model bias towards some doma ins ensuring that the global model continuously aligns with an unbiased model. T he proposed method is generic and can be combined with other existing FL methods to enhance fairness. Comprehensive experiments on Digits and Office-Caltech dem onstrate the high fairness and performance of our method.
****
Activity-Biometrics: Person Identification from Daily Activities
Shehreen Azad, Yogesh Singh Rawat; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2024, pp. 287-296
In this work we study a novel problem which focuses on person identification whi le performing daily activities. Learning biometric features from RGB videos is c hallenging due to spatio-temporal complexity and presence of appearance biases s uch as clothing color and background. We propose ABNet a novel framework which l everages disentanglement of biometric and non-biometric features to perform effe ctive person identification from daily activities. ABNet relies on a bias-less t eacher to learn biometric features from RGB videos and explicitly disentangle no n-biometric features with the help of biometric distortion. In addition ABNet al so exploits activity prior for biometrics which is enabled by joint biometric an d activity learning. We perform comprehensive evaluation of the proposed approac h across five different datasets which are derived from existing activity recogn ition benchmarks. Furthermore we extensively compare ABNet with existing works i n person identification and demonstrate its effectiveness for activity-based bio metrics across all five datasets. The code and dataset can be accessed at: https ://github.com/sacrcv/Activity-Biometrics/
****
Z*: Zero-shot Style Transfer via Attention Reweighting
Yingying Deng, Xiangyu He, Fan Tang, Weiming Dong; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6934-6944
Despite the remarkable progress in image style transfer formulating style in the context of art is inherently subjective and challenging. In contrast to existin g methods this study shows that vanilla diffusion models can directly extract st yle information and seamlessly integrate the generative prior into the content i mage without retraining. Specifically we adopt dual denoising paths to represent content/style references in latent space and then guide the content image denoi sing process with style latent codes. We further reveal that the cross-attention mechanism in latent diffusion models tends to blend the content and style image s resulting in stylized outputs that deviate from the original content image. To overcome this limitation we introduce a cross-attention reweighting strategy. T hrough theoretical analysis and experiments we demonstrate the effectiveness and

superiority of the diffusion-based zero-shot style transfer via attention reweighting Z-STAR.
****
HIG: Hierarchical Interlacement Graph Approach to Scene Graph Generation in Video Understanding
Trong-Thuan Nguyen, Pha Nguyen, Khoa Luu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18384-18394
Visual interactivity understanding within visual scenes presents a significant challenge in computer vision. Existing methods focus on complex interactivities while leveraging a simple relationship model. These methods however struggle with a diversity of appearance situation position interaction and relation in videos. This limitation hinders the ability to fully comprehend the interplay within the complex visual dynamics of subjects. In this paper we delve into interactivities understanding within visual content by deriving scene graph representations from dense interactivities among humans and objects. To achieve this goal we first present a new dataset containing Appearance-Situation-Position-Interaction-Relation predicates named ASPIRe offering an extensive collection of videos marked by a wide range of interactivities. Then we propose a new approach named Hierarchical Interlacement Graph (HIG) which leverages a unified layer and graph within a hierarchical structure to provide deep insights into scene changes across five distinct tasks. Our approach demonstrates superior performance to other methods through extensive experiments conducted in various scenarios.
****
OOSTraj: Out-of-Sight Trajectory Prediction With Vision-Positioning Denoising
Haichao Zhang, Yi Xu, Hongsheng Lu, Takayuki Shimizu, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14802-14811
Trajectory prediction is fundamental in computer vision and autonomous driving particularly for understanding pedestrian behavior and enabling proactive decision-making. Existing approaches in this field often assume precise and complete observational data neglecting the challenges associated with out-of-view objects and the noise inherent in sensor data due to limited camera range physical obstructions and the absence of ground truth for denoised sensor data. Such oversights are critical safety concerns as they can result in missing essential non-visible objects. To bridge this gap we present a novel method for out-of-sight trajectory prediction that leverages a vision-positioning technique. Our approach denoises noisy sensor observations in an unsupervised manner and precisely maps sensor-based trajectories of out-of-sight objects into visual trajectories. This method has demonstrated state-of-the-art performance in out-of-sight noisy sensor trajectory denoising and prediction on the Vi-Fi and JRDB datasets. By enhancing trajectory prediction accuracy and addressing the challenges of out-of-sight objects our work significantly contributes to improving the safety and reliability of autonomous driving in complex environments. Our work represents the first initiative towards Out-Of-Sight Trajectory prediction (OOSTraj) setting a new benchmark for future research.
****
FADES: Fair Disentanglement with Sensitive Relevance
Taeuk Jang, Xiaoqian Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12067-12076
Learning fair representation in deep learning is essential to mitigate discriminatory outcomes and enhance trustworthiness. However previous research has been commonly established on inappropriate assumptions prone to unrealistic counterfactuals and performance degradation. Although some proposed alternative approaches such as employing correlation-aware causal graphs or proxies for mutual information these methods are less practical and not applicable in general. In this work we propose FAir DisEntanglement with Sensitive relevance (FADES) a novel approach that leverages conditional mutual information from the information theory perspective to address these challenges. We employ sensitive relevant code to direct correlated information between target labels and sensitive attributes by imposing conditional independence allowing better separation of the features of inte

rest in the latent space. Utilizing an intuitive disentangling approach FADES co
nsistently achieves superior performance and fairness both quantitatively and qu
alitatively with its straightforward structure. Specifically the proposed method
 outperforms existing works in downstream classification and counterfactual gene
rations on various benchmarks.
****
Learning Continuous 3D Words for Text-to-Image Generation
Ta-Ying Cheng, Matheus Gadelha, Thibault Groueix, Matthew Fisher, Radomir Mech,
Andrew Markham, Niki Trigoni; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2024, pp. 6753-6762
Current controls over diffusion models (e.g. through text or ControlNet) for ima
ge generation fall short in recognizing abstract continuous attributes like illu
mination direction or non-rigid shape change. In this paper we present an approa
ch for allowing users of text-to-image models to have fine-grained control of se
veral attributes in an image. We do this by engineering special sets of input to
kens that can be transformed in a continuous manner we call them Continuous 3D W
ords. These attributes can for example be represented as sliders and applied joi
ntly with text prompts for fine-grained control over image generation. Given onl
y a single mesh and a rendering engine we show that our approach can be adopted
to provide continuous user control over several 3D-aware attributes including ti
me-of-day illumination bird wing orientation dollyzoom effect and object poses.
Our method is capable of conditioning image creation with multiple Continuous 3D
 Words and text descriptions simultaneously while adding no overhead to the gene
rative process.
****
MarkovGen: Structured Prediction for Efficient Text-to-Image Generation
Sadeep Jayasumana, Daniel Glasner, Srikumar Ramalingam, Andreas Veit, Ayan Chakr
abarti, Sanjiv Kumar; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2024, pp. 9316-9325
Modern text-to-image generation models produce high-quality images that are both
 photorealistic and faithful to the text prompts. However this quality comes at
significant computational cost: nearly all of these models are iterative and req
uire running sampling multiple times with large models. This iterative process i
s needed to ensure that different regions of the image are not only aligned with
 the text prompt but also compatible with each other. In this work we propose a
light-weight approach to achieving this compatibility between different regions
of an image using a Markov Random Field (MRF) model. We demonstrate the effectiv
eness of this method on top of the latent token-based Muse text-to-image model.
The MRF richly encodes the compatibility among image tokens at different spatial
 locations to improve quality and significantly reduce the required number of Mu
se sampling steps. Inference with the MRF is significantly cheaper and its param
eters can be quickly learned through back-propagation by modeling MRF inference
as a differentiable neural-network layer. Our full model MarkovGen uses this pro
posed MRF model to both speed up Muse by 1.5xand produce higher quality images b
y decreasing undesirable image artifacts.
****
Self-Supervised Class-Agnostic Motion Prediction with Spatial and Temporal Consi
stency Regularizations
Kewei Wang, Yizheng Wu, Jun Cen, Zhiyu Pan, Xingyi Li, Zhe Wang, Zhiguo Cao, Guo
sheng Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2024, pp. 14638-14647
The perception of motion behavior in a dynamic environment holds significant imp
ortance for autonomous driving systems wherein class-agnostic motion prediction
methods directly predict the motion of the entire point cloud. While most existi
ng methods rely on fully-supervised learning the manual labeling of point cloud
data is laborious and time-consuming. Therefore several annotation-efficient met
hods have been proposed to address this challenge. Although effective these meth
ods rely on weak annotations or additional multi-modal data like images and the
potential benefits inherent in the point cloud sequence are still underexplored.
 To this end we explore the feasibility of self-supervised motion prediction wit

h only unlabeled LiDAR point clouds. Initially we employ an optimal transport solver to establish coarse correspondences between current and future point clouds as the coarse pseudo motion labels. Training models directly using such coarse labels leads to noticeable spatial and temporal prediction inconsistencies. To mitigate these issues we introduce three simple spatial and temporal regularization losses which facilitate the self-supervised training process effectively. Experimental results demonstrate the significant superiority of our approach over the state-of-the-art self-supervised methods. Code will be available.
****
HashPoint: Accelerated Point Searching and Sampling for Neural Rendering
Jiahao Ma, Miaomiao Liu, David Ahmedt-Aristizabal, Chuong Nguyen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4462-4472
In this paper we address the problem of efficient point searching and sampling for volume neural rendering. Within this realm two typical approaches are employed: rasterization and ray tracing. The rasterization-based methods enable real-time rendering at the cost of increased memory and lower fidelity. In contrast the ray-tracing-based methods yield superior quality but demand longer rendering time. We solve this problem by our HashPoint method combining these two strategies leveraging rasterization for efficient point searching and sampling and ray marching for rendering. Our method optimizes point searching by rasterizing points within the camera's view organizing them in a hash table and facilitating rapid searches. Notably we accelerate the rendering process by adaptive sampling on the primary surface encountered by the ray. Our approach yields substantial speed-up for a range of state-of-the-art ray-tracing-based methods maintaining equivalent or superior accuracy across synthetic and real test datasets. The code will be available at https://jiahao-ma.github.io/hashpoint/
****
MFP: Making Full Use of Probability Maps for Interactive Image Segmentation
Chaewon Lee, Seon-Ho Lee, Chang-Su Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4051-4059
In recent interactive segmentation algorithms previous probability maps are used as network input to help predictions in the current segmentation round. However despite the utilization of previous masks useful information contained in the probability maps is not well propagated to the current predictions. In this paper to overcome this limitation we propose a novel and effective algorithm for click-based interactive image segmentation called MFP which attempts to make full use of probability maps. We first modulate previous probability maps to enhance their representations of user-specified objects. Then we feed the modulated probability maps as additional input to the segmentation network. We implement the proposed MFP algorithm based on the ResNet-34 HRNet-18 and ViT-B backbones and assess the performance extensively on various datasets. It is demonstrated that MFP meaningfully outperforms the existing algorithms using identical backbones. The source codes are available at https://github.com/cwlee00/MFP.
****
CAT: Exploiting Inter-Class Dynamics for Domain Adaptive Object Detection
Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, Robby T. Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16541-16550
Domain adaptive object detection aims to adapt detection models to domains where annotated data is unavailable. Existing methods have been proposed to address the domain gap using the semi-supervised student-teacher framework. However a fundamental issue arises from the class imbalance in the labelled training set which can result in inaccurate pseudo-labels. The relationship between classes especially where one class is a majority and the other minority has a large impact on class bias. We propose Class-Aware Teacher (CAT) to address the class bias issue in the domain adaptation setting. In our work we approximate the class relationships with our Inter-Class Relation module (ICRm) and exploit it to reduce the bias within the model. In this way we are able to apply augmentations to highly related classes both inter- and intra-domain to boost the performance of minorit

y classes while having minimal impact on majority classes. We further reduce the bias by implementing a class-relation weight to our classification loss. Experiments conducted on various datasets and ablation studies show that our method is able to address the class bias in the domain adaptation setting. On the Cityscapes ? Foggy Cityscapes dataset we attained a 52.5 mAP a substantial improvement over the 51.2 mAP achieved by the state-of-the-art method.
****
StyLitGAN: Image-Based Relighting via Latent Control
Anand Bhattad, James Soole, D.A. Forsyth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4231-4240
We describe a novel method StyLitGAN for relighting and resurfacing images in the absence of labeled data. StyLitGAN generates images with realistic lighting effects including cast shadows soft shadows inter-reflections and glossy effects without the need for paired or CGI data. StyLitGAN uses an intrinsic image method to decompose an image followed by a search of the latent space of a pretrained StyleGAN to identify a set of directions. By prompting the model to fix one component (e.g. albedo) and vary another (e.g. shading) we generate relighted images by adding the identified directions to the latent style codes. Quantitative metrics of change in albedo and lighting diversity allow us to choose effective directions using a forward selection process. Qualitative evaluation confirms the effectiveness of our method.
****
An Empirical Study of Scaling Law for Scene Text Recognition
Miao Rang, Zhenni Bi, Chuanjian Liu, Yunhe Wang, Kai Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15619-15629
The laws of model size data volume computation and model performance have been extensively studied in the field of Natural Language Processing (NLP). However the scaling laws in Scene Text Recognition (STR) have not yet been investigated. To address this we conducted comprehensive studies that involved examining the correlations between performance and the scale of models data volume and computation in the field of text recognition. Conclusively the study demonstrates smooth power laws between performance and model size as well as training data volume when other influencing factors are held constant. Additionally we have constructed a large-scale dataset called REBU-Syn which comprises 6 million real samples and 18 million synthetic samples. Based on our scaling law and new dataset we have successfully trained a scene text recognition model achieving a new state-of-the-art on 6 common test benchmarks with a top-1 average accuracy of 97.42%. The models and dataset are publicly available at \href https://github.com/large-ocr-model/large-ocr-model.github.io large-ocr-model.github.io .
****
Text2Loc: 3D Point Cloud Localization from Natural Language
Yan Xia, Letian Shi, Zifeng Ding, Joao F. Henriques, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14958-14967
We tackle the problem of 3D point cloud localization based on a few natural linguistic descriptions and introduce a novel neural network Text2Loc that fully interprets the semantic relationship between points and text. Text2Loc follows a coarse-to-fine localization pipeline: text-submap global place recognition followed by fine localization. In global place recognition relational dynamics among each textual hint are captured in a hierarchical transformer with max-pooling (HTM) whereas a balance between positive and negative pairs is maintained using text-submap contrastive learning. Moreover we propose a novel matching-free fine localization method to further refine the location predictions which completely removes the need for complicated text-instance matching and is lighter faster and more accurate than previous methods. Extensive experiments show that Text2Loc improves the localization accuracy by up to 2x over the state-of-the-art on the KITTI360Pose dataset. Our project page is publicly available at: https: //yan-xia.github.io/projects/text2loc/.
****

SVDinsTN: A Tensor Network Paradigm for Efficient Structure Search from Regulari zed Modeling Perspective
Yu-Bang Zheng, Xi-Le Zhao, Junhua Zeng, Chao Li, Qibin Zhao, Heng-Chao Li, Ting-Zhu Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26254-26263
Tensor network (TN) representation is a powerful technique for computer vision a nd machine learning. TN structure search (TN-SS) aims to search for a customized structure to achieve a compact representation which is a challenging NP-hard pr oblem. Recent "sampling-evaluation"-based methods require sampling an extensive collection of structures and evaluating them one by one resulting in prohibitive ly high computational costs. To address this issue we propose a novel TN paradig m named SVD-inspired TN decomposition (SVDinsTN) which allows us to efficiently solve the TN-SS problem from a regularized modeling perspective eliminating the repeated structure evaluations. To be specific by inserting a diagonal factor fo r each edge of the fully-connected TN SVDinsTN allows us to calculate TN cores a nd diagonal factors simultaneously with the factor sparsity revealing a compact TN structure. In theory we prove a convergence guarantee for the proposed method . Experimental results demonstrate that the proposed method achieves approximate ly 100 1000 times acceleration compared to the state-of-the-art TN-SS methods wh ile maintaining a comparable level of representation ability.
****
Decomposing Disease Descriptions for Enhanced Pathology Detection: A Multi-Aspec t Vision-Language Pre-training Framework
Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang , Zhibin Liao, Qi Wu, Minh-Son To, Johan W. Verjans; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11492-1 1501
Medical vision language pre-training (VLP) has emerged as a frontier of research enabling zero-shot pathological recognition by comparing the query image with t he textual descriptions for each disease. Due to the complex semantics of biomed ical texts current methods struggle to align medical images with key pathologica l findings in unstructured reports. This leads to the misalignment with the targ et disease's textual representation. In this paper we introduce a novel VLP fram ework designed to dissect disease descriptions into their fundamental aspects le veraging prior knowledge about the visual manifestations of pathologies. This is achieved by consulting a large language model and medical experts. Integrating a Transformer module our approach aligns an input image with the diverse element s of a disease generating aspect-centric image representations. By consolidating the matches from each aspect we improve the compatibility between an image and its associated disease. Additionally capitalizing on the aspect-oriented represe ntations we present a dual-head Transformer tailored to process known and unknow n diseases optimizing the comprehensive detection efficacy. Conducting experimen ts on seven downstream datasets ours improves the accuracy of recent methods by up to 8.56% and 17.26% for seen and unseen categories respectively. Our code is released at https://github.com/HieuPhan33/MAVL.
****
MoMask: Generative Masked Modeling of 3D Human Motions
Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, Li Cheng; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1900-1910
We introduce MoMask a novel masked modeling framework for text-driven 3D human m otion generation. In MoMask a hierarchical quantization scheme is employed to re present human motion as multi-layer discrete motion tokens with high-fidelity de tails. Starting at the base layer with a sequence of motion tokens obtained by v ector quantization the residual tokens of increasing orders are derived and stor ed at the subsequent layers of the hierarchy. This is consequently followed by t wo distinct bidirectional transformers. For the base-layer motion tokens a Maske d Transformer is designated to predict randomly masked motion tokens conditioned on text input at training stage. During generation (i.e. inference) stage start ing from an empty sequence our Masked Transformer iteratively fills up the missi

ng tokens; Subsequently a Residual Transformer learns to progressively predict the next-layer tokens based on the results from current layer. Extensive experiments demonstrate that MoMask outperforms the state-of-art methods on the text-to-motion generation task with an FID of 0.045 (vs e.g. 0.141 of T2M-GPT) on the HumanML3D dataset and 0.228 (vs 0.514) on KIT-ML respectively. MoMask can also be seamlessly applied in related tasks without further model fine-tuning such as text-guided temporal inpainting.
****
Inverse Rendering of Glossy Objects via the Neural Plenoptic Function and Radiance Fields
Haoyuan Wang, Wenbo Hu, Lei Zhu, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19999-20008
Inverse rendering aims at recovering both geometry and materials of objects. It provides a more compatible reconstruction for conventional rendering engines compared with the neural radiance fields (NeRFs). On the other hand existing NeRF-based inverse rendering methods cannot handle glossy objects with local light interactions well as they typically oversimplify the illumination as a 2D environmental map which assumes infinite lights only. Observing the superiority of NeRFs in recovering radiance fields we propose a novel 5D Neural Plenoptic Function (NeP) based on NeRFs and ray tracing such that more accurate lighting-object interactions can be formulated via the rendering equation. We also design a material-aware cone sampling strategy to efficiently integrate lights inside the BRDF lobes with the help of pre-filtered radiance fields. Our method has two stages: the geometry of the target object and the pre-filtered environmental radiance fields are reconstructed in the first stage and materials of the target object are estimated in the second stage with the proposed NeP and material-aware cone sampling strategy. Extensive experiments on the proposed real-world and synthetic datasets demonstrate that our method can reconstruct high-fidelity geometry/materials of challenging glossy objects with complex lighting interactions from nearby objects. Project webpage: https://whyy.site/paper/nep
****
Split to Merge: Unifying Separated Modalities for Unsupervised Domain Adaptation
Xinyao Li, Yuke Li, Zhekai Du, Fengling Li, Ke Lu, Jingjing Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23364-23374
Large vision-language models (VLMs) like CLIP have demonstrated good zero-shot learning performance in the unsupervised domain adaptation task. Yet most transfer approaches for VLMs focus on either the language or visual branches overlooking the nuanced interplay between both modalities. In this work we introduce a Unified Modality Separation (UniMoS) framework for unsupervised domain adaptation. Leveraging insights from modality gap studies we craft a nimble modality separation network that distinctly disentangles CLIP's features into language-associated and vision-associated components. Our proposed Modality-Ensemble Training (MET) method fosters the exchange of modality-agnostic information while maintaining modality-specific nuances. We align features across domains using a modality discriminator. Comprehensive evaluations on three benchmarks reveal our approach sets a new state-of-the-art with minimal computational costs. Code: https://github.com/TL-UESTC/UniMoS.
****
Fitting Flats to Flats
Gabriel Dogadov, Ugo Finnendahl, Marc Alexa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5439-5447
Affine subspaces of Euclidean spaces are also referred to as flats. A standard task in computer vision or more generally in engineering and applied sciences is fitting a flat to a set of points which is commonly solved using the PCA. We generalize this technique to enable fitting a flat to a set of other flats possibly of varying dimensions based on representing the flats as squared distance fields. Compared to previous approaches such as Riemannian centers of mass in the manifold of affine Grassmannians our approach is conceptually much simpler and comp

utationally more efficient yet offers desirable properties such as respecting sy
mmetries and being equivariant to rigid transformations leading to more intuitiv
e and useful results in practice. We demonstrate these claims in a number of syn
thetic experiments and a multi-view reconstruction task of line-like objects.
****
Fusing Personal and Environmental Cues for Identification and Segmentation of Fi
rst-Person Camera Wearers in Third-Person Views
Ziwei Zhao, Yuchen Wang, Chuhua Wang; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16477-16487
As wearable cameras become more popular an important question emerges: how to id
entify camera wearers within the perspective of conventional static cameras. The
 drastic difference between first-person (egocentric) and third-person (exocentr
ic) camera views makes this a challenging task. We present PersonEnvironmentNet
(PEN) a framework designed to integrate information from both the individuals in
 the two views and geometric cues inferred from the background environment. To f
acilitate research in this direction we also present TF2023 a novel dataset comp
rising synchronized first-person and third-person views along with masks of came
ra wearers and labels associating these masks with the respective first-person v
iews. In addition we propose a novel quantitative metric designed to measure a m
odel's ability to comprehend the relationship between the two views. Our experim
ents reveal that PEN outperforms existing methods. The code and dataset are avai
lable at https://github.com/ziweizhao1993/PEN.
****
Coupled Laplacian Eigenmaps for Locally-Aware 3D Rigid Point Cloud Matching
Matteo Bastico, Etienne Decencière, Laurent Corté, Yannick Tillier, David Ryckel
ynck; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2024, pp. 3447-3458
Point cloud matching a crucial technique in computer vision medical and robotics
 fields is primarily concerned with finding correspondences between pairs of poi
nt clouds or voxels. In some practical scenarios emphasizing local differences i
s crucial for accurately identifying a correct match thereby enhancing the overa
ll robustness and reliability of the matching process. Commonly used shape descr
iptors have several limitations and often fail to provide meaningful local insig
hts about the paired geometries. In this work we propose a new technique based o
n graph Laplacian eigenmaps to match point clouds by taking into account fine lo
cal structures. To deal with the order and sign ambiguity of Laplacian eigenmaps
 we introduce a new operator called Coupled Laplacian that allows to easily gene
rate aligned eigenspaces for multiple registered geometries. We show that the si
milarity between those aligned high-dimensional spaces provides a locally meanin
gful score to match shapes. We firstly evaluate the performance of the proposed
technique in a point-wise manner focusing on the task of object anomaly localiza
tion on the MVTec 3D-AD dataset. Additionally we define a new medical task calle
d automatic Bone Side Estimation (BSE) which we address through a global similar
ity score derived from coupled eigenspaces. In order to test it we propose a ben
chmark collecting bone surface structures from various public datasets. Our matc
hing technique based on Coupled Laplacian outperforms other methods by reaching
an impressive accuracy on both tasks.
****
Overcoming Generic Knowledge Loss with Selective Parameter Update
Wenxuan Zhang, Paul Janson, Rahaf Aljundi, Mohamed Elhoseiny; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 24046-24056
Foundation models encompass an extensive knowledge base and offer remarkable tra
nsferability. However this knowledge becomes outdated or insufficient over time.
 The challenge lies in continuously updating foundation models to accommodate no
vel information while retaining their original capabilities. Leveraging the fact
 that foundation models have initial knowledge on various tasks and domains we p
ropose a novel approach that instead of updating all parameters equally localize
s the updates to a sparse set of parameters relevant to the task being learned.
We strike a balance between efficiency and new task performance while maintainin

g the transferability and generalizability of foundation models. We extensively evaluate our method on foundational vision-language models with a diverse spectrum of continual learning tasks. Our method achieves improvements on the accuracy of the newly learned tasks up to 7% while preserving the pretraining knowledge with a negligible decrease of 0.9% on a representative control set accuracy.
****
Desigen: A Pipeline for Controllable Design Template Generation
Haohan Weng, Danqing Huang, Yu Qiao, Zheng Hu, Chin-Yew Lin, Tong Zhang, C. L. Philip Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12721-12732
Templates serve as a good starting point to implement a design (e.g. banner slide) but it takes great effort from designers to manually create. In this paper we present Desigen an automatic template creation pipeline which generates background images as well as harmonious layout elements over the background. Different from natural images a background image should preserve enough non-salient space for the overlaying layout elements. To equip existing advanced diffusion-based models with stronger spatial control we propose two simple but effective techniques to constrain the saliency distribution and reduce the attention weight in desired regions during the background generation process. Then conditioned on the background we synthesize the layout with a Transformer-based autoregressive generator. To achieve a more harmonious composition we propose an iterative inference strategy to adjust the synthesized background and layout in multiple rounds. We constructed a design dataset with more than 40k advertisement banners to verify our approach. Extensive experiments demonstrate that the proposed pipeline generates high-quality templates comparable to human designers. More than a single-page design we further show an application of presentation generation that outputs a set of theme-consistent slides. The data and code are available at https://whaohan.github.io/desigen.
****
Diff-BGM: A Diffusion Model for Video Background Music Generation
Sizhe Li, Yiming Qin, Minghang Zheng, Xin Jin, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27348-27357
When editing a video a piece of attractive background music is indispensable. However video background music generation tasks face several challenges for example the lack of suitable training datasets and the difficulties in flexibly controlling the music generation process and sequentially aligning the video and music. In this work we first propose a high-quality music-video dataset BGM909 with detailed annotation and shot detection to provide multi-modal information about the video and music. We then present evaluation metrics to assess music quality including music diversity and alignment between music and video with retrieval precision metrics. Finally we propose the Diff-BGM framework to automatically generate the background music for a given video which uses different signals to control different aspects of the music during the generation process i.e. uses dynamic video features to control music rhythm and semantic features to control the melody and atmosphere. We propose to align the video and music sequentially by introducing a segment-aware cross-attention layer. Experiments verify the effectiveness of our proposed method. The code and models are available at https://github.com/sizhelee/Diff-BGM.
****
Looking Similar Sounding Different: Leveraging Counterfactual Cross-Modal Pairs for Audiovisual Representation Learning
Nikhil Singh, Chih-Wei Wu, Iroro Orife, Mahdi Kalayeh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26907-26918
Audiovisual representation learning typically relies on the correspondence between sight and sound. However there are often multiple audio tracks that can correspond with a visual scene. Consider for example different conversations on the same crowded street. The effect of such counterfactual pairs on audiovisual representation learning has not been previously explored. To investigate this we use

dubbed versions of movies and television shows to augment cross-modal contrastiv
e learning. Our approach learns to represent alternate audio tracks differing on
ly in speech similarly to the same video. Our results from a comprehensive set o
f experiments investigating different training strategies show this general appr
oach improves performance on a range of downstream auditory and audiovisual task
s without majorly affecting linguistic task performance overall. These findings
highlight the importance of considering speech variation when learning scene-lev
el audiovisual correspondences and suggest that dubbed audio can be a useful aug
mentation technique for training audiovisual models toward more robust performan
ce on diverse downstream tasks.
****
Multi-criteria Token Fusion with One-step-ahead Attention for Efficient Vision T
ransformers
Sanghyeok Lee, Joonmyung Choi, Hyunwoo J. Kim; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15741-15750
Vision Transformer (ViT) has emerged as a prominent backbone for computer vision
. For more efficient ViTs recent works lessen the quadratic cost of the self-att
ention layer by pruning or fusing the redundant tokens. However these works face
d the speed-accuracy trade-off caused by the loss of information. Here we argue
that token fusion needs to consider diverse relations between tokens to minimize
 information loss. In this paper we propose a Multi-criteria Token Fusion (MCTF)
 that gradually fuses the tokens based on multi-criteria (i.e. similarity inform
ativeness and size of fused tokens). Further we utilize the one-step-ahead atten
tion which is the improved approach to capture the informativeness of the tokens
. By training the model equipped with MCTF using a token reduction consistency w
e achieve the best speed-accuracy trade-off in the image classification (ImageNe
t1K). Experimental results prove that MCTF consistently surpasses the previous r
eduction methods with and without training. Specifically DeiT-T and DeiT-S with
MCTF reduce FLOPs by about 44% while improving the performance (+0.5% and +0.3%)
 over the base model respectively. We also demonstrate the applicability of MCTF
 in various Vision Transformers (e.g. T2T-ViT LV-ViT) achieving at least 31% spe
edup without performance degradation. Code is available at https://github.com/ml
vlab/MCTF.
****
Towards HDR and HFR Video from Rolling-Mixed-Bit Spikings
Yakun Chang, Yeliduosi Xiaokaiti, Yujia Liu, Bin Fan, Zhaojun Huang, Tiejun Huan
g, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2024, pp. 25117-25127
The spiking cameras offer the benefits of high dynamic range (HDR) high temporal
 resolution and low data redundancy. However reconstructing HDR videos in high-s
peed conditions using single-bit spikings presents challenges due to the limited
 bit depth. Increasing the bit depth of the spikings is advantageous for boostin
g HDR performance but the readout efficiency will be decreased which is unfavora
ble for achieving a high frame rate (HFR) video. To address these challenges we
propose a readout mechanism to obtain rolling-mixed-bit (RMB) spikings which inv
olves interleaving multi-bit spikings within the single-bit spikings in a rollin
g manner thereby combining the characteristics of high bit depth and efficient r
eadout. Furthermore we introduce RMB-Net for reconstructing HDR and HFR videos.
RMB-Net comprises a cross-bit attention block for fusing mixed-bit spikings and
a cross-time attention block for achieving temporal fusion. Extensive experiment
s conducted on synthetic and real-synthetic data demonstrate the superiority of
our method. For instance pure 3-bit spikings result in 3 times of data volume wh
ereas our method achieves comparable performance with less than 2% increase in d
ata volume.
****
Scaling Up Video Summarization Pretraining with Large Language Models
Dawit Mureja Argaw, Seunghyun Yoon, Fabian Caba Heilbron, Hanieh Deilamsalehy, T
rung Bui, Zhaowen Wang, Franck Dernoncourt, Joon Son Chung; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
8332-8341

Long-form video content constitutes a significant portion of internet traffic making automated video summarization an essential research problem. However existing video summarization datasets are notably limited in their size constraining the effectiveness of state-of-the-art methods for generalization. Our work aims to overcome this limitation by capitalizing on the abundance of long-form videos with dense speech-to-video alignment and the remarkable capabilities of recent large language models (LLMs) in summarizing long text. We introduce an automated and scalable pipeline for generating a large-scale video summarization dataset using LLMs as Oracle summarizers. By leveraging the generated dataset we analyze the limitations of existing approaches and propose a new video summarization model that effectively addresses them. To facilitate further research in the field our work also presents a new benchmark dataset that contains 1200 long videos each with high-quality summaries annotated by professionals. Extensive experiments clearly indicate that our proposed approach sets a new state-of-the-art in video summarization across several benchmarks.
****
Continuous Optical Zooming: A Benchmark for Arbitrary-Scale Image Super-Resolution in Real World
Huiyuan Fu, Fei Peng, Xianwei Li, Yejun Li, Xin Wang, Huadong Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3035-3044
Most current arbitrary-scale image super-resolution (SR) methods has commonly relied on simulated data generated by simple synthetic degradation models (e.g. bicubic downsampling) at continuous various scales thereby falling short in capturing the complex degradation of real-world images. This limitation hinders the visual quality of these methods when applied to real-world images. To address this issue we propose the Continuous Optical Zooming dataset (COZ) by constructing an automatic imaging system to collect images at fine-grained various focal lengths within a specific range and providing strict image pair alignment. The COZ dataset serves as a benchmark to provide real-world data for training and testing arbitrary-scale SR models. To enhance the model's robustness against real-world image degradation we propose a Local Mix Implicit network (LMI) based on the MLP-mixer architecture and meta-learning which directly learns the local texture information by simultaneously mixing features and coordinates of multiple independent points. The extensive experiments demonstrate the superior performance of the arbitrary-scale SR models trained on the COZ dataset compared to models trained on simulated data. Our LMI model exhibits the superior effectiveness compared to other models. This study is of great significance in developing more efficient algorithms and improving the performance of arbitrary-scale image SR methods in practical applications. Our dataset and codes are available at https://github.com/pf0607/COZ.
****
Sharingan: A Transformer Architecture for Multi-Person Gaze Following
Samy Tafasca, Anshul Gupta, Jean-Marc Odobez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2008-2017
Gaze is a powerful form of non-verbal communication that humans develop from an early age. As such modeling this behavior is an important task that can benefit a broad set of application domains ranging from robotics to sociology. In particular the gaze following task in computer vision is defined as the prediction of the 2D pixel coordinates where a person in the image is looking. Previous attempts in this area have primarily centered on CNN-based architectures but they have been constrained by the need to process one person at a time which proves to be highly inefficient. In this paper we introduce a novel and effective multi-person transformer-based architecture for gaze prediction. While there exist prior works using transformers for multi-person gaze prediction they use a fixed set of learnable embeddings to decode both the person and its gaze target which requires a matching step afterward to link the predictions with the annotations. Thus it is difficult to quantitatively evaluate these methods reliably with the available benchmarks or integrate them into a larger human behavior understanding system. Instead we are the first to propose a multi-person transformer-based archit

ecture that maintains the original task formulation and ensures control over the people fed as input. Our main contribution lies in encoding the person-specific information into a single controlled token to be processed alongside image tokens and using its output for prediction based on a novel multiscale decoding mechanism. Our new architecture achieves state-of-the-art results on the GazeFollow VideoAttentionTarget and ChildPlay datasets and outperforms comparable multi-person architectures with a notable margin. Our code checkpoints and data extractions will be made publicly available soon.
****
ViewFusion: Towards Multi-View Consistency via Interpolated Denoising
Xianghui Yang, Yan Zuo, Sameera Ramasinghe, Loris Bazzani, Gil Avraham, Anton van den Hengel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9870-9880
Novel-view synthesis through diffusion models has demonstrated remarkable potential for generating diverse and high-quality images. Yet the independent process of image generation in these prevailing methods leads to challenges in maintaining multiple-view consistency. To address this we introduce ViewFusion a novel training-free algorithm that can be seamlessly integrated into existing pre-trained diffusion models. Our approach adopts an auto-regressive method that implicitly leverages previously generated views as context for the next view generation ensuring robust multi-view consistency during the novel-view generation process. Through a diffusion process that fuses known-view information via interpolated denoising our framework successfully extends single-view conditioned models to work in multiple-view conditional settings without any additional fine-tuning. Extensive experimental results demonstrate the effectiveness of ViewFusion in generating consistent and detailed novel views.
****
SketchINR: A First Look into Sketches as Implicit Neural Representations
Hmrishav Bandyopadhyay, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Tao Xiang, Timothy Hospedales, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12565-12574
We propose SketchINR to advance the representation of vector sketches with implicit neural models. A variable length vector sketch is compressed into a latent space of fixed dimension that implicitly encodes the underlying shape as a function of time and strokes. The learned function predicts the xy point coordinates in a sketch at each time and stroke. Despite its simplicity SketchINR outperforms existing representations at multiple tasks: (i) Encoding an entire sketch dataset into a fixed size latent vector SketchINR gives 60x and 10x data compression over raster and vector sketches respectively. (ii) SketchINR's auto-decoder provides a much higher-fidelity representation than other learned vector sketch representations and is uniquely able to scale to complex vector sketches such as FS-COCO. (iii) SketchINR supports parallelisation that can decode/render 100x faster than other learned vector representations such as SketchRNN. (iv) SketchINR for the first time emulates the human ability to reproduce a sketch with varying abstraction in terms of number and complexity of strokes. As a first look at implicit sketches SketchINR's compact high-fidelity representation will support future work in modelling long and complex sketches.
****
Open-Vocabulary Segmentation with Semantic-Assisted Calibration
Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, Yansong Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3491-3500
This paper studies open-vocabulary segmentation (OVS) through calibrating in-vocabulary and domain-biased embedding space with generalized contextual prior of CLIP. As the core of open-vocabulary understanding alignment of visual content with the semantics of unbounded text has become the bottleneck of this field. To address this challenge recent works propose to utilize CLIP as an additional classifier and aggregate model predictions with CLIP classification results. Despite their remarkable progress performance of OVS methods in relevant scenarios is still unsatisfactory compared with supervised counterparts. We attribute this to

the in-vocabulary embedding and domain-biased CLIP prediction. To this end we pr esent a Semantic-assisted CAlibration Network (SCAN). In SCAN we incorporate gen eralized semantic prior of CLIP into proposal embedding to avoid collapsing on k nown categories. Besides a contextual shift strategy is applied to mitigate the lack of global context and unnatural background noise. With above designs SCAN a chieves state-of-the-art performance on all popular open-vocabulary segmentation benchmarks. Furthermore we also focus on the problem of existing evaluation sys tem that ignores semantic duplication across categories and propose a new metric called Semantic-Guided IoU (SG-IoU).
****

MatchU: Matching Unseen Objects for 6D Pose Estimation from RGB-D Images
Junwen Huang, Hao Yu, Kuan-Ting Yu, Nassir Navab, Slobodan Ilic, Benjamin Busam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2024, pp. 10095-10105
Recent learning methods for object pose estimation require resource-intensive tr aining for each individual object instance or category hampering their scalabili ty in real applications when confronted with previously unseen objects. In this paper we propose MatchU a Fuse-Describe-Match strategy for 6D pose estimation fr om RGB-D images. MatchU is a generic approach that fuses 2D texture and 3D geome tric cues for 6D pose prediction of unseen objects. We rely on learning geometri c 3D descriptors that are rotation-invariant by design. By encoding pose-agnosti c geometry the learned descriptors naturally generalize to unseen objects and ca pture symmetries. To tackle ambiguous associations using 3D geometry only we fus e additional RGB information into our descriptor. This is achieved through a nov el attention-based mechanism that fuses cross-modal information together with a matching loss that leverages the latent space learned from RGB data to guide the descriptor learning process. Extensive experiments reveal the generalizability of both the RGB-D fusion strategy as well as the descriptor efficacy. Benefiting from the novel designs MatchU surpasses all existing methods by a significant m argin in terms of both accuracy and speed even without the requirement of expens ive re-training or rendering.
****

Towards a Perceptual Evaluation Framework for Lighting Estimation
Justine Giroux, Mohammad Reza Karimi Dastjerdi, Yannick Hold-Geoffroy, Javier Va zquez-Corral, Jean-François Lalonde; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2024, pp. 4410-4419
Progress in lighting estimation is tracked by computing existing image quality a ssessment (IQA) metrics on images from standard datasets. While this may appear to be a reasonable approach we demonstrate that doing so does not correlate to h uman preference when the estimated lighting is used to relight a virtual scene i nto a real photograph. To study this we design a controlled psychophysical exper iment where human observers must choose their preference amongst rendered scenes lit using a set of lighting estimation algorithms selected from the recent lite rature and use it to analyse how these algorithms perform according to human per ception. Then we demonstrate that none of the most popular IQA metrics from the literature taken individually correctly represent human perception. Finally we s how that by learning a combination of existing IQA metrics we can more accuratel y represent human preference. This provides a new perceptual framework to help e valuate future lighting estimation algorithms. To encourage future research all (anonymised) perceptual data and code are available at https://lvsn.github.io/Pe rceptionMetric/.
****

Bridging the Synthetic-to-Authentic Gap: Distortion-Guided Unsupervised Domain A daptation for Blind Image Quality Assessment
Aobo Li, Jinjian Wu, Yongxu Liu, Leida Li; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28422-28431
The annotation of blind image quality assessment (BIQA) is labor-intensive and t ime-consuming especially for authentic images. Training on synthetic data is exp ected to be beneficial but synthetically trained models often suffer from poor g eneralization in real domains due to domain gaps. In this work we make a key obs

ervation that introducing more distortion types in the synthetic dataset may not improve or even be harmful to generalizing authentic image quality assessment. To solve this challenge we propose distortion-guided unsupervised domain adaptation for BIQA (DGQA) a novel framework that leverages adaptive multi-domain selection via prior knowledge from distortion to match the data distribution between the source domains and the target domain thereby reducing negative transfer from the outlier source domains. Extensive experiments on two cross-domain settings (synthetic distortion to authentic distortion and synthetic distortion to algorithmic distortion) have demonstrated the effectiveness of our proposed DGQA. Besides DGQA is orthogonal to existing model-based BIQA methods and can be used in combination with such models to improve performance with less training data.
****
Coherent Temporal Synthesis for Incremental Action Segmentation
Guodong Ding, Hans Golong, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28485-28494
Data replay is a successful incremental learning technique for images. It prevents catastrophic forgetting by keeping a reservoir of previous data original or synthesized to ensure the model retains past knowledge while adapting to novel concepts. However its application in the video domain is rudimentary as it simply stores frame exemplars for action recognition. This paper presents the first exploration of video data replay techniques for incremental action segmentation focusing on action temporal modeling. We propose a Temporally Coherent Action (TCA) model which represents actions using a generative model instead of storing individual frames. The integration of a conditioning variable that captures temporal coherence allows our model to understand the evolution of action features over time. Therefore action segments generated by TCA for replay are diverse and temporally coherent. In a 10-task incremental setup on the Breakfast dataset our approach achieves significant increases in accuracy for up to 22% compared to the baselines.
****
HiFi4G: High-Fidelity Human Performance Rendering via Compact Gaussian Splatting
Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, Lan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19734-19745
We have recently seen tremendous progress in photo-real human modeling and rendering. Yet efficiently rendering realistic human performance and integrating it into the rasterization pipeline remains challenging. In this paper we present HiFi4G an explicit and compact Gaussian-based approach for high-fidelity human performance rendering from dense footage. Our core intuition is to marry the 3D Gaussian representation with non-rigid tracking achieving a compact and compression-friendly representation. We first propose a dual-graph mechanism to obtain motion priors with a coarse deformation graph for effective initialization and a fine-grained Gaussian graph to enforce subsequent constraints. Then we utilize a 4D Gaussian optimization scheme with adaptive spatial-temporal regularizers to effectively balance the non-rigid prior and Gaussian updating. We also present a companion compression scheme with residual compensation for immersive experiences on various platforms. It achieves a substantial compression rate of approximately 25 times with less than 2MB of storage per frame. Extensive experiments demonstrate the effectiveness of our approach which significantly outperforms existing approaches in terms of optimization speed rendering quality and storage overhead.
****
G-FARS: Gradient-Field-based Auto-Regressive Sampling for 3D Part Grouping
Junfeng Cheng, Tania Stathaki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27652-27661
This paper proposes a novel task named "3D part grouping". Suppose there is a mixed set containing scattered parts from various shapes. This task requires algorithms to find out every possible combination among all the parts. To address this challenge we propose the so called Gradient Field-based Auto-Regressive Sampling framework (G-FARS) tailored specifically for the 3D part grouping task. In ou

r framework we design a gradient-field-based selection graph neural network (GNN) to learn the gradients of a log conditional probability density in terms of part selection where the condition is the given mixed part set. This innovative approach implemented through the gradient-field-based selection GNN effectively captures complex relationships among all the parts in the input. Upon completion of the training process our framework becomes capable of autonomously grouping 3D parts by iteratively selecting them from the mixed part set leveraging the knowledge acquired by the trained gradient-field-based selection GNN. Our code is available at: https://github.com/J-F-Cheng/G-FARS-3DPartGrouping.
****
Towards High-fidelity Artistic Image Vectorization via Texture-Encapsulated Shape Parameterization
Ye Chen, Bingbing Ni, Jinfan Liu, Xiaoyang Huang, Xuanhong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15877-15886
We develop a novel vectorized image representation scheme accommodating both shape/geometry and texture in a decoupled way particularly tailored for reconstruction and editing tasks of artistic/design images such as Emojis and Cliparts. In the heart of this representation is a set of sparsely and unevenly located 2D control points. On one hand these points constitute a collection of parametric/vectorized geometric primitives (e.g. curves and closed shapes) describing the shape characteristics of the target image. On the other hand local texture codes in terms of implicit neural network parameters are spatially distributed into each control point yielding local coordinate-to-RGB mappings within the anchored region of each control point. In the meantime a zero-shot learning algorithm is developed to decompose an arbitrary raster image into the above representation for the sake of high-fidelity image vectorization with convenient editing ability. Extensive experiments on a series of image vectorization and editing tasks well demonstrate the high accuracy offered by our proposed method with a significantly higher image compression ratio over prior art.
****
On Exact Inversion of DPM-Solvers
Seongmin Hong, Kyeonghyun Lee, Suh Yoon Jeon, Hyewon Bae, Se Young Chun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7069-7078
Diffusion probabilistic models (DPMs) are a key component in modern generative models. DPM-solvers have achieved reduced latency and enhanced quality significantly but have posed challenges to find the exact inverse (i.e. finding the initial noise from the given image). Here we investigate the exact inversions for DPM-solvers and propose algorithms to perform them when samples are generated by the first-order as well as higher-order DPM-solvers. For each explicit denoising step in DPM-solvers we formulated the inversions using implicit methods such as gradient descent or forward step method to ensure the robustness to large classifier-free guidance unlike the prior approach using fixed-point iteration. Experimental results demonstrated that our proposed exact inversion methods significantly reduced the error of both image and noise reconstructions greatly enhanced the ability to distinguish invisible watermarks and well prevented unintended background changes consistently during image editing.
****
EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything
Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, Raghuraman Krishnamoorthi, Vikas Chandra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16111-16121
Segment Anything Model (SAM) has emerged as a powerful tool for numerous vision applications. A key component that drives the impressive performance for zero-shot transfer and high versatility is a super large Transformer model trained on the extensive high-quality SA-1B dataset. While beneficial the huge computation cost of SAM model has limited its applications to wider real-world applications. To address this limitation we propose EfficientSAMs light-weight SAM models that

exhibits decent performance with largely reduced complexity. Our idea is based on leveraging masked image pretraining SAMI which learns to reconstruct features from SAM image encoder for effective visual representation learning. Further we take SAMI-pretrained light-weight image encoders and mask decoder to build EfficientSAMs and finetune the models on SA-1B for segment anything task. We perform evaluations on multiple vision tasks including image classification object detection instance segmentation and semantic segmentation and find that our proposed pretraining method SAMI consistently outperforms other masked image pretraining methods. On segment anything task such as zero-shot instance segmentation our EfficientSAMs with SAMI-pretrained lightweight image encoders perform favorably with a significant gain (e.g. 4 AP on COCO/LVIS) over other fast SAM models. Our EfficientSAM code and models are available at https://github.com/yformer/EfficientSAM.
****
ChatScene: Knowledge-Enabled Safety-Critical Scenario Generation for Autonomous Vehicles
Jiawei Zhang, Chejian Xu, Bo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15459-15469
We present ChatScene a Large Language Model (LLM)-based agent that leverages the capabilities of LLMs to generate safety-critical scenarios for autonomous vehicles. Given unstructured language instructions the agent first generates textually described traffic scenarios using LLMs. These scenario descriptions are subsequently broken down into several sub-descriptions for specified details such as behaviors and locations of vehicles. The agent then distinctively transforms the textually described sub-scenarios into domain-specific languages which then generate actual code for prediction and control in simulators facilitating the creation of diverse and complex scenarios within the CARLA simulation environment. A key part of our agent is a comprehensive knowledge retrieval component which efficiently translates specific textual descriptions into corresponding domain-specific code snippets by training a knowledge database containing the scenario description and code pairs. Extensive experimental results underscore the efficacy of ChatScene in improving the safety of autonomous vehicles. For instance the scenarios generated by ChatScene show a 15% increase in collision rates compared to state-of-the-art baselines when tested against different reinforcement learning-based ego vehicles. Furthermore we show that by using our generated safety-critical scenarios to fine-tune different RL-based autonomous driving models they can achieve a 9% reduction in collision rates surpassing current SOTA methods. ChatScene effectively bridges the gap between textual descriptions of traffic scenarios and practical CARLA simulations providing a unified way to conveniently generate safety-critical scenarios for safety testing and improvement for AVs.
****
CAMEL: CAusal Motion Enhancement Tailored for Lifting Text-driven Video Editing
Guiwei Zhang, Tianyu Zhang, Guanglin Niu, Zichang Tan, Yalong Bai, Qing Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9079-9088
Text-driven video editing poses significant challenges in exhibiting flicker-free visual continuity while preserving the inherent motion patterns of original videos. Existing methods operate under a paradigm where motion and appearance are intricately intertwined. This coupling leads to the network either over-fitting appearance content -- failing to capture motion patterns -- or focusing on motion patterns at the expense of content generalization to diverse textual scenarios. Inspired by the pivotal role of wavelet transform in dissecting video sequences we propose CAusal Motion Enhancement tailored for Lifting text-driven video editing (CAMEL) a novel technique with two core designs. First we introduce motion prompts designed to summarize motion concepts from video templates through direct optimization. The optimized prompts are purposefully integrated into latent representations of diffusion models to enhance the motion fidelity of generated results. Second to enhance motion coherence and extend the generalization of appearance content to creative textual prompts we propose the causal motion-enhanced attention mechanism. This mechanism is implemented in tandem with a novel causa

l motion filter synergistically enhancing the motion coherence of disentangled high-frequency components and concurrently preserving the generalization of appearance content across various textual scenarios. Extensive experimental results show the superior performance of CAMEL.
****
Teeth-SEG: An Efficient Instance Segmentation Framework for Orthodontic Treatment based on Multi-Scale Aggregation and Anthropic Prior Knowledge
Bo Zou, Shaofeng Wang, Hao Liu, Gaoyue Sun, Yajie Wang, FeiFei Zuo, Chengbin Quan, Youjian Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11601-11610
Teeth localization segmentation and labeling in 2D images have great potential in modern dentistry to enhance dental diagnostics treatment planning and population-based studies on oral health. However general instance segmentation frameworks are incompetent due to 1) the subtle differences between some teeth' shapes (e.g. maxillary first premolar and second premolar) 2) the teeth's position and shape variation across subjects and 3) the presence of abnormalities in the dentition (e.g. caries and edentulism). To address these problems we propose a ViT-based framework named TeethSEG which consists of stacked Multi-Scale Aggregation (MSA) blocks and an Anthropic Prior Knowledge (APK) layer. Specifically to compose the two modules we design 1) a unique permutation-based upscaler to ensure high efficiency while establishing clear segmentation boundaries with 2) multi-head self/cross-gating layers to emphasize particular semantics meanwhile maintaining the divergence between token embeddings. Besides we collect 3) the first open-sourced intraoral image dataset IO150K which comprises over 150k intraoral photos and all photos are annotated by orthodontists using a human-machine hybrid algorithm. Experiments on IO150K demonstrate that our TeethSEG outperforms the state-of-the-art segmentation models on dental image segmentation.
****
FocSAM: Delving Deeply into Focused Objects in Segmenting Anything
You Huang, Zongyu Lan, Liujuan Cao, Xianming Lin, Shengchuan Zhang, Guannan Jiang, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3120-3130
The Segment Anything Model (SAM) marks a notable milestone in segmentation models highlighted by its robust zero-shot capabilities and ability to handle diverse prompts. SAM follows a pipeline that separates interactive segmentation into image preprocessing through a large encoder and interactive inference via a lightweight decoder ensuring efficient real-time performance. However SAM faces stability issues in challenging samples upon this pipeline. These issues arise from two main factors. Firstly the image preprocessing disables SAM to dynamically use image-level zoom-in strategies to refocus on the target object during interaction. Secondly the lightweight decoder struggles to sufficiently integrate interactive information with image embeddings. To address these two limitations we propose FocSAM with a pipeline redesigned on two pivotal aspects. First we propose Dynamic Window Multi-head Self-Attention (Dwin-MSA) to dynamically refocus SAM's image embeddings on the target object. Dwin-MSA localizes attention computations around the target object enhancing object-related embeddings with minimal computational overhead. Second we propose Pixel-wise Dynamic ReLU (P-DyReLU) to enable sufficient integration of interactive information from a few initial clicks that have significant impacts on the overall segmentation results. Experimentally FocSAM augments SAM's interactive segmentation performance to match the existing state-of-the-art method in segmentation quality requiring only about 5.6% of this method's inference time on CPUs. Code is available at https://github.com/YouHuang67/focsam.
****
DMR: Decomposed Multi-Modality Representations for Frames and Events Fusion in Visual Reinforcement Learning
Haoran Xu, Peixi Peng, Guang Tan, Yuan Li, Xinhai Xu, Yonghong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26508-26518
We explore visual reinforcement learning (RL) using two complementary visual mod

alities: frame-based RGB camera and event-based Dynamic Vision Sensor (DVS). Existing multi-modality visual RL methods often encounter challenges in effectively extracting task-relevant information from multiple modalities while suppressing the increased noise only using indirect reward signals instead of pixel-level supervision. To tackle this we propose a Decomposed Multi-Modality Representation (DMR) framework for visual RL. It explicitly decomposes the inputs into three distinct components: combined task-relevant features (co-features) RGB-specific noise and DVS-specific noise. The co-features represent the full information from both modalities that is relevant to the RL task; the two noise components each constrained by a data reconstruction loss to avoid information leak are contrasted with the co-features to maximize their difference. Extensive experiments demonstrate that by explicitly separating the different types of information our approach achieves substantially improved policy performance compared to state-of-the-art approaches.
****
DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models
Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, Karthik Nandakumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27621-27630
Recently a number of image-mixing-based augmentation techniques have been introduced to improve the generalization of deep neural networks. In these techniques two or more randomly selected natural images are mixed together to generate an augmented image. Such methods may not only omit important portions of the input images but also introduce label ambiguities by mixing images across labels resulting in misleading supervisory signals. To address these limitations we propose DIFFUSEMIX a novel data augmentation technique that leverages a diffusion model to reshape training images supervised by our bespoke conditional prompts. First concatenation of a partial natural image and its generated counterpart is obtained which helps in avoiding the generation of unrealistic images or label ambiguities. Then to enhance resilience against adversarial attacks and improves safety measures a randomly selected structural pattern from a set of fractal images is blended into the concatenated image to form the final augmented image for training. Our empirical results on seven different datasets reveal that DIFFUSEMIX achieves superior performance compared to existing state- of-the-art methods on tasks including general classification fine-grained classification fine-tuning data scarcity and adversarial robustness.
****
PRDP: Proximal Reward Difference Prediction for Large-Scale Reward Finetuning of Diffusion Models
Fei Deng, Qifei Wang, Wei Wei, Tingbo Hou, Matthias Grundmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7423-7433
Reward finetuning has emerged as a promising approach to aligning foundation models with downstream objectives. Remarkable success has been achieved in the language domain by using reinforcement learning (RL) to maximize rewards that reflect human preference. However in the vision domain existing RL-based reward finetuning methods are limited by their instability in large-scale training rendering them incapable of generalizing to complex unseen prompts. In this paper we propose Proximal Reward Difference Prediction (PRDP) enabling stable black-box reward finetuning for diffusion models for the first time on large-scale prompt datasets with over 100K prompts. Our key innovation is the Reward Difference Prediction (RDP) objective that has the same optimal solution as the RL objective while enjoying better training stability. Specifically the RDP objective is a supervised regression objective that tasks the diffusion model with predicting the reward difference of generated image pairs from their denoising trajectories. We theoretically prove that the diffusion model that obtains perfect reward difference prediction is exactly the maximizer of the RL objective. We further develop an online algorithm with proximal updates to stably optimize the RDP objective. In experiments we demonstrate that PRDP can match the reward maximization ability of well-established RL-based methods in small-scale training. Furthermore through l

arge-scale training on text prompts from the Human Preference Dataset v2 and the Pick-a-Pic v1 dataset PRDP achieves superior generation quality on a diverse set of complex unseen prompts whereas RL-based methods completely fail.
****

FREE: Faster and Better Data-Free Meta-Learning

Yongxian Wei, Zixuan Hu, Zhenyi Wang, Li Shen, Chun Yuan, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23273-23282

Data-Free Meta-Learning (DFML) aims to extract knowledge from a collection of pre-trained models without requiring the original data presenting practical benefits in contexts constrained by data privacy concerns. Current DFML methods primarily focus on the data recovery from these pre-trained models. However they suffer from slow recovery speed and overlook gaps inherent in heterogeneous pre-trained models. In response to these challenges we introduce the Faster and Better Data-Free Meta-Learning (FREE) framework which contains: (i) a meta-generator for rapidly recovering training tasks from pre-trained models; and (ii) a meta-learner for generalizing to new unseen tasks. Specifically within the module Faster Inversion via Meta-Generator each pre-trained model is perceived as a distinct task. The meta-generator can rapidly adapt to a specific task in just five steps significantly accelerating the data recovery. Furthermore we propose Better Generalization via Meta-Learner and introduce an implicit gradient alignment algorithm to optimize the meta-learner. This is achieved as aligned gradient directions alleviate potential conflicts among tasks from heterogeneous pre-trained models. Empirical experiments on multiple benchmarks affirm the superiority of our approach marking a notable speed-up (20x) and performance enhancement (1.42%   4.78%) in comparison to the state-of-the-art.
****

Bayesian Diffusion Models for 3D Shape Reconstruction

Haiyang Xu, Yu Lei, Zeyuan Chen, Xiang Zhang, Yue Zhao, Yilin Wang, Zhuowen Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10628-10638

We present Bayesian Diffusion Models (BDM) a prediction algorithm that performs effective Bayesian inference by tightly coupling the top-down (prior) information with the bottom-up (data-driven) procedure via joint diffusion processes. We demonstrate the application of BDM on the 3D shape reconstruction task. Compared to standard deep learning data-driven approaches relying on supervised data our BDM can bring in rich prior information trained in an unsupervised manner to improve the bottom-up 3D reconstruction. As opposed to the traditional Bayesian frameworks where explicitly learned prior and data-driven distributions are required for gradient computation and combination BDM performs a seamless fusion of the two via coupled diffusion processes with learned gradient computation networks. The specialty of our Bayesian Diffusion Models (BDM) lies in its capability to engage the active and effective information exchange and fusion of the top-down and bottom-up processes where each itself is a diffusion process. We demonstrate state-of-the-art results on both synthetic and real-world benchmarks for 3D shape reconstruction. Project link: https://mlpc-ucsd.github.io/BDM
****

Task-Customized Mixture of Adapters for General Image Fusion

Pengfei Zhu, Yang Sun, Bing Cao, Qinghua Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7099-7108

General image fusion aims at integrating important information from multi-source images. However due to the significant cross-task gap the respective fusion mechanism varies considerably in practice resulting in limited performance across subtasks. To handle this problem we propose a novel task-customized mixture of adapters (TC-MoA) for general image fusion adaptively prompting various fusion tasks in a unified model. We borrow the insight from the mixture of experts (MoE) taking the experts as efficient tuning adapters to prompt a pre-trained foundation model. These adapters are shared across different tasks and constrained by mutual information regularization ensuring compatibility with different tasks while complementarity for multi-source images. The task-specific routing networks cus

tomize these adapters to extract task-specific information from different source
s with dynamic dominant intensity performing adaptive visual feature prompt fusi
on. Notably our TC-MoA controls the dominant intensity bias for different fusion
 tasks successfully unifying multiple fusion tasks in a single model. Extensive
experiments show that TC-MoA outperforms the competing approaches in learning co
mmonalities while retaining compatibility for general image fusion (multi-modal
multi-exposure and multi-focus) and also demonstrating striking controllability
on more generalization experiments. The code is available at https://github.com/
YangSun22/TC-MoA.
****
Bi-SSC: Geometric-Semantic Bidirectional Fusion for Camera-based 3D Semantic Sce
ne Completion
Yujie Xue, Ruihui Li, Fan Wu, Zhuo Tang, Kenli Li, Mingxing Duan; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 20124-20134
Camera-based Semantic Scene Completion (SSC) is to infer the full geometry of ob
jects and scenes from only 2D images. The task is particularly challenging for t
hose invisible areas due to the inherent occlusions and lighting ambiguity. Exis
ting works ignore the information missing or ambiguous in those shaded and occlu
ded areas resulting in distorted geometric prediction. To address this issue we
propose a novel method Bi-SSC bidirectional geometric semantic fusion for camera
-based 3D semantic scene completion. The key insight is to use the neighboring s
tructure of objects in the image and the spatial differences from different pers
pectives to compensate for the lack of information in occluded areas. Specifical
ly we introduce a spatial sensory fusion module with multiple association attent
ion to improve semantic correlation in geometric distributions. This module work
s within single view and across stereo views to achieve global spatial consisten
cy. Experimental results demonstrate that Bi-SSC outperforms state-of-the-art ca
mera-based methods on SemanticKITTI particularly excelling in those invisible an
d shaded areas.
****
CrossKD: Cross-Head Knowledge Distillation for Object Detection
Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, Qibin Hou; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 16520-16530
Knowledge Distillation (KD) has been validated as an effective model compression
 technique for learning compact object detectors. Existing state-of-the-art KD m
ethods for object detection are mostly based on feature imitation. In this paper
 we present a general and effective prediction mimicking distillation scheme cal
led CrossKD which delivers the intermediate features of the student's detection
head to the teacher's detection head. The resulting cross-head predictions are t
hen forced to mimic the teacher's predictions. This manner relieves the student'
s head from receiving contradictory supervision signals from the annotations and
 the teacher's predictions greatly improving the student's detection performance
. Moreover as mimicking the teacher's predictions is the target of KD CrossKD of
fers more task-oriented information in contrast with feature imitation. On MS CO
CO with only prediction mimicking losses applied our CrossKD boosts the average
precision of GFL ResNet-50 with 1x training schedule from 40.2 to 43.7 outperfor
ming all existing KD methods. In addition our method also works well when distil
ling detectors with heterogeneous backbones.
****
Bi-level Learning of Task-Specific Decoders for Joint Registration and One-Shot
Medical Image Segmentation
Xin Fan, Xiaolin Wang, Jiaxin Gao, Jia Wang, Zhongxuan Luo, Risheng Liu; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 11726-11735
One-shot medical image segmentation (MIS) aims to cope with the expensive time-c
onsuming and inherent human bias annotations. One prevalent method to address on
e-shot MIS is joint registration and segmentation (JRS) with a shared encoder wh
ich mainly explores the voxel-wise correspondence between the labeled data and u

nlabeled data for better segmentation. However this method omits underlying conn ections between task-specific decoders for segmentation and registration leading to unstable training. In this paper we propose a novel Bi-level Learning of Tas k-Specific Decoders for one-shot MIS employing a pretrained fixed shared encoder that is proved to be more quickly adapted to brand-new datasets than existing J RS without fixed shared encoder paradigm. To be more specific we introduce a bi-level optimization training strategy considering registration as a major objecti ve and segmentation as a learnable constraint by leveraging inter-task coupling dependencies. Furthermore we design an appearance conformity constraint strategy that learns the backward transformations generating the fake labeled data used to perform data augmentation instead of the labeled image to avoid performance d egradation caused by inconsistent styles between unlabeled data and labeled data in previous methods. Extensive experiments on the brain MRI task across ABIDE A DNI and PPMI datasets demonstrate that the proposed Bi-JROS outperforms state-of -the-art one-shot MIS methods for both segmentation and registration tasks. The code will be available at https://github.com/Coradlut/Bi-JROS.
****
Parameter Efficient Self-Supervised Geospatial Domain Adaptation
Linus Scheibenreif, Michael Mommert, Damian Borth; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27841-278 51
As large-scale foundation models become publicly available for different domains efficiently adapting them to individual downstream applications and additional data modalities has turned into a central challenge. For example foundation mode ls for geospatial and satellite remote sensing applications are commonly trained on large optical RGB or multi-spectral datasets although data from a wide varie ty of heterogeneous sensors are available in the remote sensing domain. This lea ds to significant discrepancies between pre-training and downstream target data distributions for many important applications. Fine-tuning large foundation mode ls to bridge that gap incurs high computational cost and can be infeasible when target datasets are small. In this paper we address the question of how large pr e-trained foundational transformer models can be efficiently adapted to downstre am remote sensing tasks involving different data modalities or limited dataset s ize. We present a self-supervised adaptation method that boosts downstream linea r evaluation accuracy of different foundation models by 4-6% (absolute) across 8 remote sensing datasets while outperforming full fine-tuning when training only 1-2% of the model parameters. Our method significantly improves label efficienc y and increases few-shot accuracy by 6-10% on different datasets.
****
Defense without Forgetting: Continual Adversarial Defense with Anisotropic & Iso tropic Pseudo Replay
Yuhang Zhou, Zhongyun Hua; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2024, pp. 24263-24272
Deep neural networks have demonstrated susceptibility to adversarial attacks. Ad versarial defense techniques often focus on one-shot setting to maintain robustn ess against attack. However new attacks can emerge in sequences in real-world de ployment scenarios. As a result it is crucial for a defense model to constantly adapt to new attacks but the adaptation process can lead to catastrophic forgett ing of previously defended against attacks. In this paper we discuss for the fir st time the concept of continual adversarial defense under a sequence of attacks and propose a lifelong defense baseline called Anisotropic & Isotropic Replay (AIR) which offers three advantages: (1) Isotropic replay ensures model consisten cy in the neighborhood distribution of new data indirectly aligning the output p reference between old and new tasks. (2) Anisotropic replay enables the model to learn a compromise data manifold with fresh mixed semantics for further replay constraints and potential future attacks. (3) A straightforward regularizer miti gates the 'plasticity-stability' trade-off by aligning model output between new and old tasks. Experiment results demonstrate that AIR can approximate or even e xceed the empirical performance upper bounds achieved by Joint Training.
****

EscherNet: A Generative Model for Scalable View Synthesis
Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9503-9513
We introduce EscherNet a multi-view conditioned diffusion model for view synthesis. EscherNet learns implicit and generative 3D representations coupled with a specialised camera positional encoding allowing precise and continuous relative control of the camera transformation between an arbitrary number of reference and target views. EscherNet offers exceptional generality flexibility and scalability in view synthesis --- it can generate more than 100 consistent target views simultaneously on a single consumer-grade GPU despite being trained with a fixed number of 3 reference views to 3 target views. As a result EscherNet not only addresses zero-shot novel view synthesis but also naturally unifies single- and multi-image 3D reconstruction combining these diverse tasks into a single cohesive framework. Our extensive experiments demonstrate that EscherNet achieves state-of-the-art performance in multiple benchmarks even when compared to methods specifically tailored for each individual problem. This remarkable versatility opens up new directions for designing scalable neural architectures for 3D vision. Project page: https://kxhit.github.io/EscherNet.
****
MeaCap: Memory-Augmented Zero-shot Image Captioning
Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, Zhengjue Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14100-14110
Zero-shot image captioning (IC) without well-paired image-text data can be categorized into two main types: training-free and text-only-training methods. While both types integrate pre-trained vision-language models such as CLIP for image-text similarity evaluation and a pre-trained language model (LM) for caption generation their distinction lies in the utilization of textual corpus for LM training. Despite achieving promising performance on certain metrics existing methods commonly suffer from drawbacks. Training-free methods often generate hallucinations whereas text-only-training methods may lack generalization capability. To address these challenges we propose a novel Memory-Augmented zero-shot image Captioning framework (MeaCap). This framework equipped with a textual memory incorporates a retrieve-then-filter module to extract key concepts highly relevant to the image. By leveraging our proposed memory-augmented visual-related fusion score within a keywords-to-sentence LM MeaCap generates concept-centered captions that exhibit high consistency with the image with reduced hallucinations and enriched world knowledge. MeaCap achieves state-of-the-art performance across various zero-shot IC settings. Our code is publicly available at https://github.com/joeyz0z/MeaCap.
****
Artist-Friendly Relightable and Animatable Neural Heads
Yingyan Xu, Prashanth Chandran, Sebastian Weiss, Markus Gross, Gaspard Zoss, Derek Bradley; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2457-2467
An increasingly common approach for creating photo-realistic digital avatars is through the use of volumetric neural fields. The original neural radiance field (NeRF) allowed for impressive novel view synthesis of static heads when trained on a set of multi-view images and follow up methods showed that these neural representations can be extended to dynamic avatars. Recently new variants also surpassed the usual drawback of baked-in illumination in neural representations showing that static neural avatars can be relit in any environment. In this work we simultaneously tackle both the motion and illumination problem proposing a new method for relightable and animatable neural heads. Our method builds on a proven dynamic avatar approach based on a mixture of volumetric primitives combined with a recently-proposed lightweight hardware setup for relightable neural fields and includes a novel architecture that allows relighting dynamic neural avatars performing unseen expressions in any environment even with nearfield illumination and viewpoints.

****
Elite360D: Towards Efficient 360 Depth Estimation via Semantic- and Distance-Awa re Bi-Projection Fusion
Hao Ai, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9926-9935
360 depth estimation has recently received great attention for 3D reconstruction owing to its omnidirectional field of view (FoV). Recent approaches are predomi nantly focused on cross-projection fusion with geometry-based re-projection: the y fuse 360 images with equirectangular projection (ERP) and another projection t ype e.g. cubemap projection to estimate depth with the ERP format. However these methods suffer from 1) limited local receptive fields making it hardly possible to capture large FoV scenes and 2) prohibitive computational cost caused by the complex cross-projection fusion module design. In this paper we propose Elite36 0D a novel framework that inputs the ERP image and icosahedron projection (ICOSA P) point set which is undistorted and spatially continuous. Elite360D is superio r in its capacity in learning a representation from a local-with-global perspect ive. With a flexible ERP image encoder it includes an ICOSAP point encoder and a Bi-projection Bi-attention Fusion (B2F) module (totally 1M parameters). Specif ically the ERP image encoder can take various perspective image-trained backbone s (e.g. ResNet Transformer) to extract local features. The point encoder extract s the global features from the ICOSAP. Then the B2F module captures the semantic - and distance-aware dependencies between each pixel of the ERP feature and the entire ICOSAP feature set. Without specific backbone design and obvious computat ional cost increase Elite360D outperforms the prior arts on several benchmark da tasets.
****
From Feature to Gaze: A Generalizable Replacement of Linear Layer for Gaze Estim ation
Yiwei Bao, Feng Lu; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2024, pp. 1409-1418
Deep-learning-based gaze estimation approaches often suffer from notable perform ance degradation in unseen target domains. One of the primary reasons is that th e Fully Connected layer is highly prone to overfitting when mapping the high-dim ensional image feature to 3D gaze. In this paper we propose Analytical Gaze Gene ralization framework (AGG) to improve the generalization ability of gaze estimat ion models without touching target domain data. The AGG consists of two modules the Geodesic Projection Module (GPM) and the Sphere-Oriented Training (SOT). GPM is a generalizable replacement of FC layer which projects high-dimensional imag e features to 3D space analytically to extract the principle components of gaze. Then we propose Sphere-Oriented Training (SOT) to incorporate the GPM into the training process and further improve cross-domain performances. Experimental res ults demonstrate that the AGG effectively alleviate the overfitting problem and consistently improves the cross-domain gaze estimation accuracy in 12 cross-doma in settings without requiring any target domain data. The insight from the Analy tical Gaze Generalization framework has the potential to benefit other regressio n tasks with physical meanings.
****
Curriculum Point Prompting for Weakly-Supervised Referring Image Segmentation
Qiyuan Dai, Sibei Yang; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2024, pp. 13711-13722
Referring image segmentation (RIS) aims to precisely segment referents in images through corresponding natural language expressions yet relying on cost-intensiv e mask annotations. Weakly supervised RIS thus learns from image-text pairs to p ixel-level semantics which is challenging for segmenting fine-grained masks. A n atural approach to enhancing segmentation precision is to empower weakly supervi sed RIS with the image segmentation foundation model SAM. Nevertheless we observ e that simply integrating SAM yields limited benefits and can even lead to perfo rmance regression due to the inevitable noise issues and challenges in excessive focus on object parts. In this paper we present an innovative framework Point P rompTing (PPT) incorporated with the proposed multi-source curriculum learning s

trategy to address these challenges. Specifically the core of PPT is a point generator that not only harnesses CLIP's text-image alignment capability and SAM's powerful mask generation ability but also generates negative point prompts to address the noisy and excessive focus issues inherently and effectively. In addition we introduce a curriculum learning strategy with object-centric images to help PPT gradually learn from simpler yet precise semantic alignment to more complex RIS. Experiments demonstrate that our PPT significantly and consistently outperforms prior weakly supervised techniques on mIoU by 11.34% 14.14% and 6.97% across RefCOCO RefCOCO+ and G-Ref respectively.
****
EventDance: Unsupervised Source-free Cross-modal Adaptation for Event-based Object Recognition
Xu Zheng, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17448-17458
In this paper we make the first attempt at achieving the cross-modal (i.e. image-to-events) adaptation for event-based object recognition without accessing any labeled source image data owning to privacy and commercial issues. Tackling this novel problem is non-trivial due to the novelty of event cameras and the distinct modality gap between images and events. In particular as only the source model is available a hurdle is how to extract the knowledge from the source model by only using the unlabeled target event data while achieving knowledge transfer. To this end we propose a novel framework dubbed EventDance for this unsupervised source-free cross-modal adaptation problem. Importantly inspired by event-to-video reconstruction methods we propose a reconstruction-based modality bridging (RMB) module which reconstructs intensity frames from events in a self-supervised manner. This makes it possible to build up the surrogate images to extract the knowledge (i.e. labels) from the source model. We then propose a multi-representation knowledge adaptation (MKA) module that transfers the knowledge to target models learning events with multiple representation types for fully exploring the spatiotemporal information of events. The two modules connecting the source and target models are mutually updated so as to achieve the best performance. Experiments on three benchmark datasets with two adaption settings show that EventDance is on par with prior methods utilizing the source data.
****
CycleINR: Cycle Implicit Neural Representation for Arbitrary-Scale Volumetric Super-Resolution of Medical Data
Wei Fang, Yuxing Tang, Heng Guo, Mingze Yuan, Tony C. W. Mok, Ke Yan, Jiawen Yao, Xin Chen, Zaiyi Liu, Le Lu, Ling Zhang, Minfeng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11631-11641
In the realm of medical 3D data such as CT and MRI images prevalent anisotropic resolution is characterized by high intra-slice but diminished inter-slice resolution. The lowered resolution between adjacent slices poses challenges hindering optimal viewing experiences and impeding the development of robust downstream analysis algorithms. Various volumetric super-resolution algorithms aim to surmount these challenges enhancing inter-slice resolution and overall 3D medical imaging quality. However existing approaches confront inherent challenges: 1) often tailored to specific upsampling factors lacking flexibility for diverse clinical scenarios; 2) newly generated slices frequently suffer from over-smoothing degrading fine details and leading to inter-slice inconsistency. In response this study presents CycleINR a novel enhanced Implicit Neural Representation model for 3D medical data volumetric super-resolution. Leveraging the continuity of the learned implicit function the CycleINR model can achieve results with arbitrary up-sampling rates eliminating the need for separate training. Additionally we enhance the grid sampling in CycleINR with a local attention mechanism and mitigate over-smoothing by integrating cycle-consistent loss. We introduce a new metric Slice-wise Noise Level Inconsistency (SNLI) to quantitatively assess inter-slice noise level inconsistency. The effectiveness of our approach is demonstrated through image quality evaluations on an in-house dataset and a downstream task analysis on the Medical Segmentation Decathlon liver tumor dataset.

****
Boosting Image Restoration via Priors from Pre-trained Models
Xiaogang Xu, Shu Kong, Tao Hu, Zhe Liu, Hujun Bao; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2900-2909
Pre-trained models with large-scale training data such as CLIP and Stable Diffus
ion have demonstrated remarkable performance in various high-level computer visi
on tasks such as image understanding and generation from language descriptions.
Yet their potential for low-level tasks such as image restoration remains relati
vely unexplored. In this paper we explore such models to enhance image restorati
on. As off-the-shelf features (OSF) from pre-trained models do not directly serv
e image restoration we propose to learn an additional lightweight module called
Pre-Train-Guided Refinement Module (PTG-RM) to refine restoration results of a t
arget restoration network with OSF. PTG-RM consists of two components Pre-Train-
Guided Spatial-Varying Enhancement (PTG-SVE) and Pre-Train-Guided Channel-Spatia
l Attention (PTG-CSA). PTG-SVE enables optimal short- and long-range neural oper
ations while PTG-CSA enhances spatial-channel attention for restoration-related
learning. Extensive experiments demonstrate that PTG-RM with its compact size (<
1M parameters) effectively enhances restoration performance of various models ac
ross different tasks including low-light enhancement deraining deblurring and de
noising.
****
VRetouchEr: Learning Cross-frame Feature Interdependence with Imperfection Flow
for Face Retouching in Videos
Wen Xue, Le Jiang, Lianxin Xie, Si Wu, Yong Xu, Hau San Wong; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 9141-9150
Face Video Retouching is a complex task that often requires labor-intensive manu
al editing. Conventional image retouching methods perform less satisfactorily in
 terms of generalization performance and stability when applied to videos withou
t exploiting the correlation among frames. To address this issue we propose a Vi
deo Retouching transformEr to remove facial imperfections in videos which is ref
erred to as VRetouchEr. Specifically we estimate the apparent motion of imperfec
tions between two consecutive frames and the resulting displacement vectors are
used to refine the imperfection map which is synthesized from the current frame
together with the corresponding encoder features. The flow-based imperfection re
finement is critical for precise and stable retouching across frames. To leverag
e the temporal contextual information we inject the refined imperfection map int
o each transformer block for multi-frame masked attention computation such that
we can capture the interdependence between the current frame and multiple refere
nce frames. As a result the imperfection regions can be replaced with normal ski
n with high fidelity while at the same time keeping the other regions unchanged.
 Extensive experiments are performed to verify the superiority of VRetouchEr ove
r state-of-the-art image retouching methods in terms of fidelity and stability.
****
Transferable Structural Sparse Adversarial Attack Via Exact Group Sparsity Train
ing
Di Ming, Peng Ren, Yunlong Wang, Xin Feng; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24696-24705
Deep neural networks (DNNs) are vulnerable to highly transferable adversarial at
tacks. Especially many studies have shown that sparse attacks pose a significant
 threat to DNNs on account of their exceptional imperceptibility. Current sparse
 attack methods mostly limit only the magnitude and number of perturbations whil
e generally overlooking the location of the perturbations resulting in decreased
 performances on attack transferability. A subset of studies indicates that pert
urbations existing in the significant regions with rich classification-relevant
features are more effective. Leveraging this insight we introduce the structural
 sparsity constraint in the framework of generative models to limit the perturba
tion positions. To ensure that the perturbations are generated towards classific
ation-relevant regions we propose an exact group sparsity training method to lea
rn pixel-level and group-level sparsity. For purpose of improving the effectiven

ess of sparse training we further put forward masked quantization network and multi-stage optimization algorithm in the training process. Utilizing CNNs as surrogate models extensive experiments demonstrate that our method has higher transferability in image classification attack compared to state-of-the-art methods at approximately same sparsity levels. In cross-model ViT object detection and semantic segmentation attack tasks we also achieve a better attack success rate. Code is available at https://github.com/MisterRpeng/EGS-TSSA.
****
Holistic Autonomous Driving Understanding by Bird's-Eye-View Injected Multi-Modal Large Models
Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, Xiaomeng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13668-13677
The rise of multimodal large language models (MLLMs) has spurred interest in language-based driving tasks. However existing research typically focuses on limited tasks and often omits key multi-view and temporal information which is crucial for robust autonomous driving. To bridge these gaps we introduce NuInstruct a novel dataset with 91K multi-view video-QA pairs across 17 subtasks where each task demands holistic information (e.g. temporal multi-view and spatial) significantly elevating the challenge level. To obtain NuInstruct we propose a novel SQL-based method to generate instruction-response pairs automatically which is inspired by the driving logical progression of humans. We further present BEV-InMLLM an end-to-end method for efficiently deriving instruction-aware Bird's-Eye-View (BEV) features language-aligned for large language models. BEV-InMLLM integrates multi-view spatial awareness and temporal semantics to enhance MLLMs' capabilities on NuInstruct tasks. Moreover our proposed BEV injection module is a plug-and-play method for existing MLLMs. Our experiments on NuInstruct demonstrate that BEV-InMLLM significantly outperforms existing MLLMs e.g 9% improvement on various tasks. We release our NuInstruct at https://github.com/xmed-lab/NuInstruct.
****
Arbitrary-Scale Image Generation and Upsampling using Latent Diffusion Model and Implicit Neural Decoder
Jinseok Kim, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9202-9211
Super-resolution (SR) and image generation are important tasks in computer vision and are widely adopted in real-world applications. Most existing methods however generate images only at fixed-scale magnification and suffer from over-smoothing and artifacts. Additionally they do not offer enough diversity of output images nor image consistency at different scales. Most relevant work applied Implicit Neural Representation (INR) to the denoising diffusion model to obtain continuous-resolution yet diverse and high-quality SR results. Since this model operates in the image space the larger the resolution of image is produced the more memory and inference time is required and it also does not maintain scale-specific consistency. We propose a novel pipeline that can super-resolve an input image or generate from a random noise a novel image at arbitrary scales. The method consists of a pretrained auto-encoder a latent diffusion model and an implicit neural decoder and their learning strategies. The proposed method adopts diffusion processes in a latent space thus efficient yet aligned with output image space decoded by MLPs at arbitrary scales. More specifically our arbitrary-scale decoder is designed by the symmetric decoder w/o up-scaling from the pretrained auto-encoder and Local Implicit Image Function (LIIF) in series. The latent diffusion process is learnt by the denoising and the alignment losses jointly. Errors in output images are backpropagated via the fixed decoder improving the quality of output images. In the extensive experiments using multiple public benchmarks on the two tasks i.e. image super-resolution and novel image generation at arbitrary scales the proposed method outperforms relevant methods in metrics of image quality diversity and scale consistency. It is significantly better than the relevant prior-art in the inference speed and memory usage.
****
Unsupervised Occupancy Learning from Sparse Point Cloud

Amine Ouasfi, Adnane Boukhayma; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2024, pp. 21729-21739
Implicit Neural Representations have gained prominence as a powerful framework f
or capturing complex data modalities encompassing a wide range from 3D shapes to
 images and audio. Within the realm of 3D shape representation Neural Signed Dis
tance Functions (SDF) have demonstrated remarkable potential in faithfully encod
ing intricate shape geometry. However learning SDFs from 3D point clouds in the
absence of ground truth supervision remains a very challenging task. In this pap
er we propose a method to infer occupancy fields instead of SDFs as they are eas
ier to learn from sparse inputs. We leverage a margin-based uncertainty measure
to differentiably sample from the decision boundary of the occupancy function an
d supervise the sampled boundary points using the input point cloud. We further
stabilise the optimization process at the early stages of the training by biasin
g the occupancy function towards minimal entropy fields while maximizing its ent
ropy at the input point cloud. Through extensive experiments and evaluations we
illustrate the efficacy of our proposed method highlighting its capacity to impr
ove implicit shape inference with respect to baselines and the state-of-the-art
using synthetic and real data.
****
Extreme Point Supervised Instance Segmentation
Hyeonjun Lee, Sehyun Hwang, Suha Kwak; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17212-17222
This paper introduces a novel approach to learning instance segmentation using e
xtreme points i.e. the topmost leftmost bottommost and rightmost points of each
object. These points are readily available in the modern bounding box annotation
 process while offering strong clues for precise segmentation and thus allows to
 improve performance at the same annotation cost with box-supervised methods. Ou
r work considers extreme points as a part of the true instance mask and propagat
es them to identify potential foreground and background points which are all tog
ether used for training a pseudo label generator. Then pseudo labels given by th
e generator are in turn used for supervised learning of our final model. On thre
e public benchmarks our method significantly outperforms existing box-supervised
 methods further narrowing the gap with its fully supervised counterpart. In par
ticular our model generates high-quality masks when a target object is separated
 into multiple parts where previous box-supervised methods often fail.
****
3DInAction: Understanding Human Actions in 3D Point Clouds
Yizhak Ben-Shabat, Oren Shrout, Stephen Gould; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19978-19987
We propose a novel method for 3D point cloud action recognition. Understanding h
uman actions in RGB videos has been widely studied in recent years however its 3
D point cloud counterpart remains under-explored despite the clear value that 3D
 information may bring. This is mostly due to the inherent limitation of the poi
nt cloud data modality---lack of structure permutation invariance and varying nu
mber of points---which makes it difficult to learn a spatio-temporal representat
ion. To address this limitation we propose the 3DinAction pipeline that first es
timates patches moving in time (t-patches) as a key building block alongside a h
ierarchical architecture that learns an informative spatio-temporal representati
on. We show that our method achieves improved performance on existing datasets i
ncluding DFAUST and IKEA ASM. Code is publicly available at https://github.com/s
itzikbs/3dincaction
****
Cache Me if You Can: Accelerating Diffusion Models through Block Caching
Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, A
rtsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, Christian Rupprecht, Da
niel Cremers, Peter Vajda, Jialiang Wang; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6211-6220
Diffusion models have recently revolutionized the field of image synthesis due t
o their ability to generate photorealistic images. However one of the major draw
backs of diffusion models is that the image generation process is costly. A larg

e image-to-image network has to be applied many times to iteratively refine an i
mage from random noise. While many recent works propose techniques to reduce the
 number of required steps they generally treat the underlying denoising network
as a black box. In this work we investigate the behavior of the layers within th
e network and find that 1) the layers' output changes smoothly over time 2) the
layers show distinct patterns of change and 3) the change from step to step is o
ften very small. We hypothesize that many layer computations in the denoising ne
twork are redundant. Leveraging this we introduce Block Caching in which we reus
e outputs from layer blocks of previous steps to speed up inference. Furthermore
 we propose a technique to automatically determine caching schedules based on ea
ch block's changes over timesteps. In our experiments we show through FID human
evaluation and qualitative analysis that Block Caching allows to generate images
 with higher visual quality at the same computational cost. We demonstrate this
for different state-of-the-art models (LDM and EMU) and solvers (DDIM and DPM).
****
MedM2G: Unifying Medical Multi-Modal Generation via Cross-Guided Diffusion with
Visual Invariant
Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, Jian Wu; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 115
02-11512
Medical generative models acknowledged for their high-quality sample generation
ability have accelerated the fast growth of medical applications. However recent
 works concentrate on separate medical generation models for distinct medical ta
sks and are restricted to inadequate medical multi-modal knowledge constraining
medical comprehensive diagnosis. In this paper we propose MedM2G a Medical Multi
-Modal Generative framework with the key innovation to align extract and generat
e medical multi-modal within a unified model. Extending beyond single or two med
ical modalities we efficiently align medical multi-modal through the central ali
gnment approach in the unified space. Significantly our framework extracts valua
ble clinical knowledge by preserving the medical visual invariant of each imagin
g modal thereby enhancing specific medical information for multi-modal generatio
n. By conditioning the adaptive cross-guided parameters into the multi-flow diff
usion framework our model promotes flexible interactions among medical multi-mod
al for generation. MedM2G is the first medical generative model that unifies med
ical generation tasks of text-to-image image-to-text and unified generation of m
edical modalities (CT MRI X-ray). It performs 5 medical generation tasks across
10 datasets consistently outperforming various state-of-the-art works.
****
SDDGR: Stable Diffusion-based Deep Generative Replay for Class Incremental Objec
t Detection
Junsu Kim, Hoseong Cho, Jihyeon Kim, Yihalem Yimolal Tiruneh, Seungryul Baek; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 28772-28781
In the field of class incremental learning (CIL) generative replay has become in
creasingly prominent as a method to mitigate the catastrophic forgetting alongsi
de the continuous improvements in generative models. However its application in
class incremental object detection (CIOD) has been significantly limited primari
ly due to the complexities of scenes involving multiple labels. In this paper we
 propose a novel approach called stable diffusion deep generative replay (SDDGR)
 for CIOD. Our method utilizes a diffusion-based generative model with pre-train
ed text-to-image diffusion networks to generate realistic and diverse synthetic
images. SDDGR incorporates an iterative refinement strategy to produce high-qual
ity images encompassing old classes. Additionally we adopt an L2 knowledge disti
llation technique to improve the retention of prior knowledge in synthetic image
s. Furthermore our approach includes pseudo-labeling for old objects within new
task images preventing misclassification as background elements. Extensive exper
iments on the COCO 2017 dataset demonstrate that SDDGR significantly outperforms
 existing algorithms achieving a new state-of-the-art in various CIOD scenarios.
****
Neural Parametric Gaussians for Monocular Non-Rigid Object Reconstruction

Devikalyan Das, Christopher Wewer, Raza Yunus, Eddy Ilg, Jan Eric Lenssen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10715-10725
Reconstructing dynamic objects from monocular videos is a severely underconstrained and challenging problem and recent work has approached it in various directions. However owing to the ill-posed nature of this problem there has been no solution that can provide consistent high-quality novel views from camera positions that are significantly different from the training views. In this work we introduce Neural Parametric Gaussians (NPGs) to take on this challenge by imposing a two-stage approach: first we fit a low-rank neural deformation model which then is used as regularization for non-rigid reconstruction in the second stage. The first stage learns the object's deformations such that it preserves consistency in novel views. The second stage obtains high reconstruction quality by optimizing 3D Gaussians that are driven by the coarse model. To this end we introduce a local 3D Gaussian representation where temporally shared Gaussians are anchored in and deformed by local oriented volumes. The resulting combined model can be rendered as radiance fields resulting in high-quality photo-realistic reconstructions of the non-rigidly deforming objects. We demonstrate that NPGs achieve superior results compared to previous works especially in challenging scenarios with few multi-view cues.
****
Physical 3D Adversarial Attacks against Monocular Depth Estimation in Autonomous Driving
Junhao Zheng, Chenhao Lin, Jiahao Sun, Zhengyu Zhao, Qian Li, Chao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24452-24461
Deep learning-based monocular depth estimation (MDE) extensively applied in autonomous driving is known to be vulnerable to adversarial attacks. Previous physical attacks against MDE models rely on 2D adversarial patches so they only affect a small localized region in the MDE map but fail under various viewpoints. To address these limitations we propose 3D Depth Fool (3D^2Fool) the first 3D texture-based adversarial attack against MDE models. 3D^2Fool is specifically optimized to generate 3D adversarial textures agnostic to model types of vehicles and to have improved robustness in bad weather conditions such as rain and fog. Experimental results validate the superior performance of our 3D^2Fool across various scenarios including vehicles MDE models weather conditions and viewpoints. Real-world experiments with printed 3D textures on physical vehicle models further demonstrate that our 3D^2Fool can cause an MDE error of over 10 meters.
****
Adaptive Random Feature Regularization on Fine-tuning Deep Neural Networks
Shin'ya Yamaguchi, Sekitoshi Kanai, Kazuki Adachi, Daiki Chijiwa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23481-23490
While fine-tuning is a de facto standard method for training deep neural networks it still suffers from overfitting when using small target datasets. Previous methods improve fine-tuning performance by maintaining knowledge of the source datasets or introducing regularization terms such as contrastive loss. However these methods require auxiliary source information (e.g. source labels or datasets) or heavy additional computations. In this paper we propose a simple method called adaptive random feature regularization (AdaRand). AdaRand helps the feature extractors of training models to adaptively change the distribution of feature vectors for downstream classification tasks without auxiliary source information and with reasonable computation costs. To this end AdaRand minimizes the gap between feature vectors and random reference vectors that are sampled from class conditional Gaussian distributions. Furthermore AdaRand dynamically updates the conditional distribution to follow the currently updated feature extractors and balance the distance between classes in feature spaces. Our experiments show that AdaRand outperforms the other fine-tuning regularization requiring auxiliary source information and heavy computation costs.
****

PH-Net: Semi-Supervised Breast Lesion Segmentation via Patch-wise Hardness
Siyao Jiang, Huisi Wu, Junyang Chen, Qin Zhang, Jing Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11418-11427
We present a novel semi-supervised framework for breast ultrasound (BUS) image segmentation which is a very challenging task owing to (1) large scale and shape variations of breast lesions and (2) extremely ambiguous boundaries caused by massive speckle noise and artifacts in BUS images. While existing models achieved certain progress in this task we believe the main bottleneck nowadays for further improvement is that we still cannot deal with hard cases well. Our framework aims to break through this bottleneck which includes two innovative components: an adaptive patch augmentation scheme and a hard-patch contrastive learning module. We first identify hard patches by computing the average entropy of each patch and then shield hard patches to prevent them from being cropped out while performing random patch cutmix. Such a scheme is able to prevent hard regions from being inadequately trained under strong augmentation. We further develop a new hard-patch contrastive learning algorithm to direct model attention to hard regions by applying extra contrast to pixels in hard patches further improving segmentation performance on hard cases. We demonstrate the superiority of our framework to state-of-the-art approaches on two famous BUS datasets achieving better performance under different labeling conditions. The code is available at https://github.com/jjjsyyy/PH-Net.
****
Multimodal Prompt Perceiver: Empower Adaptiveness Generalizability and Fidelity for All-in-One Image Restoration
Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, Ran He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25432-25444
Despite substantial progress all-in-one image restoration (IR) grapples with persistent challenges in handling intricate real-world degradations. This paper introduces MPerceiver: a novel multimodal prompt learning approach that harnesses Stable Diffusion (SD) priors to enhance adaptiveness generalizability and fidelity for all-in-one image restoration. Specifically we develop a dual-branch module to master two types of SD prompts: textual for holistic representation and visual for multiscale detail representation. Both prompts are dynamically adjusted by degradation predictions from the CLIP image encoder enabling adaptive responses to diverse unknown degradations. Moreover a plug-in detail refinement module improves restoration fidelity via direct encoder-to-decoder information transformation. To assess our method MPerceiver is trained on 9 tasks for all-in-one IR and outperforms state-of-the-art task-specific methods across many tasks. Post multitask pre-training MPerceiver attains a generalized representation in low-level vision exhibiting remarkable zero-shot and few-shot capabilities in unseen tasks. Extensive experiments on 16 IR tasks underscore the superiority of MPerceiver in terms of adaptiveness generalizability and fidelity.
****
ExACT: Language-guided Conceptual Reasoning and Uncertainty Estimation for Event-based Action Recognition and More
Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18633-18643
Event cameras have recently been shown beneficial for practical vision tasks such as action recognition thanks to their high temporal resolution power efficiency and reduced privacy concerns. However current research is hindered by 1) the difficulty in processing events because of their prolonged duration and dynamic actions with complex and ambiguous semantics and 2) the redundant action depiction of the event frame representation with fixed stacks. We find language naturally conveys abundant semantic information rendering it stunningly superior in reducing semantic uncertainty. In light of this we propose ExACT a novel approach that for the first time tackles event-based action recognition from a cross-modal conceptualizing perspective. Our ExACT brings two technical contributions. Firstly we propose an adaptive fine-grained event (AFE) representation to adaptively

filter out the repeated events for the stationary objects while preserving dynam
ic ones. This subtly enhances the performance of ExACT without extra computation
al cost. Then we propose a conceptual reasoning-based uncertainty estimation mod
ule which simulates the recognition process to enrich the semantic representatio
n. In particular conceptual reasoning builds the temporal relation based on the
action semantics and uncertainty estimation tackles the semantic uncertainty of
actions based on the distributional representation. Experiments show that our Ex
ACT achieves superior recognition accuracy of 94.83%(+2.23%) 90.10%(+37.47%) and
67.24% on PAF HARDVS and our SeAct datasets respectively.
****
Color Shift Estimation-and-Correction for Image Enhancement
Yiyu Li, Ke Xu, Gerhard Petrus Hancke, Rynson W.H. Lau; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2538
9-25398
Images captured under sub-optimal illumination conditions may contain both over-
and under-exposures. We observe that over- and over-exposed regions display opp
osite color tone distribution shifts which may not be easily normalized in joint
modeling as they usually do not have "normal-exposed" regions/pixels as referen
ce. In this paper we propose a novel method to enhance images with both over- an
d under-exposures by learning to estimate and correct such color shifts. Specifi
cally we first derive the color feature maps of the brightened and darkened vers
ions of the input image via a UNet-based network followed by a pseudo-normal fea
ture generator to produce pseudo-normal color feature maps. We then propose a no
vel COlor Shift Estimation (COSE) module to estimate the color shifts between th
e derived brightened (or darkened) color feature maps and the pseudo-normal colo
r feature maps. The COSE module corrects the estimated color shifts of the over-
and under-exposed regions separately. We further propose a novel COlor MOdulati
on (COMO) module to modulate the separately corrected colors in the over- and un
der-exposed regions to produce the enhanced image. Comprehensive experiments sho
w that our method outperforms existing approaches.
****
Improving Visual Recognition with Hyperbolical Visual Hierarchy Mapping
Hyeongjun Kwon, Jinhyun Jang, Jin Kim, Kwonyoung Kim, Kwanghoon Sohn; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 17364-17374
Visual scenes are naturally organized in a hierarchy where a coarse semantic is
recursively comprised of several fine details. Exploring such a visual hierarchy
is crucial to recognize the complex relations of visual elements leading to a c
omprehensive scene understanding. In this paper we propose a Visual Hierarchy Ma
pper (Hi-Mapper) a novel approach for enhancing the structured understanding of
the pre-trained Deep Neural Networks (DNNs). Hi-Mapper investigates the hierarch
ical organization of the visual scene by 1) pre-defining a hierarchy tree throug
h the encapsulation of probability densities; and 2) learning the hierarchical r
elations in hyperbolic space with a novel hierarchical contrastive loss. The pre
-defined hierarchy tree recursively interacts with the visual features of the pr
e-trained DNNs through hierarchy decomposition and encoding procedures thereby e
ffectively identifying the visual hierarchy and enhancing the recognition of an
entire scene. Extensive experiments demonstrate that Hi-Mapper significantly enh
ances the representation capability of DNNs leading to an improved performance o
n various tasks including image classification and dense prediction tasks.
****
ParameterNet: Parameters Are All You Need for Large-scale Visual Pretraining of
Mobile Networks
Kai Han, Yunhe Wang, Jianyuan Guo, Enhua Wu; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15751-15761
The large-scale visual pretraining has significantly improve the performance of
large vision models. However we observe the low FLOPs pitfall that the existing
low-FLOPs models cannot benefit from large-scale pretraining. In this paper we i
ntroduce a novel design principle termed ParameterNet aimed at augmenting the nu
mber of parameters in large-scale visual pretraining models while minimizing the

increase in FLOPs. We leverage dynamic convolutions to incorporate additional parameters into the networks with only a marginal rise in FLOPs. The ParameterNet approach allows low-FLOPs networks to take advantage of large-scale visual pretraining. Furthermore we extend the ParameterNet concept to the language domain to enhance inference results while preserving inference speed. Experiments on the large-scale ImageNet-22K have shown the superiority of our ParameterNet scheme. For example ParameterNet-600M can achieve higher accuracy than the widely-used Swin Transformer (81.6% vs. 80.9%) and has much lower FLOPs (0.6G vs. 4.5G). The code will be released at https://parameternet.github.io/.
****
Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation
Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, Konrad Schindler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9492-9502
Monocular depth estimation is a fundamental computer vision task. Recovering 3D depth from a single image is geometrically ill-posed and requires scene understanding so it is not surprising that the rise of deep learning has led to a breakthrough. The impressive progress of monocular depth estimators has mirrored the growth in model capacity from relatively modest CNNs to large Transformer architectures. Still monocular depth estimators tend to struggle when presented with images with unfamiliar content and layout since their knowledge of the visual world is restricted by the data seen during training and challenged by zero-shot generalization to new domains. This motivates us to explore whether the extensive priors captured in recent generative diffusion models can enable better more generalizable depth estimation. We introduce Marigold a method for affine-invariant monocular depth estimation that is derived from Stable Diffusion and retains its rich prior knowledge. The estimator can be fine-tuned in a couple of days on a single GPU using only synthetic training data. It delivers state-of-the-art performance across a wide range of datasets including over 20% performance gains in specific cases. Project page: https://marigoldmonodepth.github.io.
****
Identifying Important Group of Pixels using Interactions
Kosuke Sumiyasu, Kazuhiko Kawamoto, Hiroshi Kera; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6017-6026
To better understand the behavior of image classifiers it is useful to visualize the contribution of individual pixels to the model prediction. In this study we propose a method MoXI(Model eXplanation by Interactions) that efficiently and accurately identifies a group of pixels with high prediction confidence. The proposed method employs game-theoretic concepts Shapley values and interactions taking into account the effects of individual pixels and the cooperative influence of pixels on model confidence. Theoretical analysis and experiments demonstrate that our method better identifies the pixels that are highly contributing to the model outputs than widely-used by Grad-CAM Attention rollout and Shapley value. While prior studies have suffered from the exponential computational cost in the computation of Shapley value and interactions we show that this can be reduced to quadratic cost for our task. The code is available at https://github.com/KosukeSumiyasu/MoXI.
****
Towards Scalable 3D Anomaly Detection and Localization: A Benchmark via 3D Anomaly Synthesis and A Self-Supervised Learning Network
Wenqiao Li, Xiaohao Xu, Yao Gu, Bozhong Zheng, Shenghua Gao, Yingna Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22207-22216
Recently 3D anomaly detection a crucial problem involving fine-grained geometry discrimination is getting more attention. However the lack of abundant real 3D anomaly data limits the scalability of current models. To enable scalable anomaly data collection we propose a 3D anomaly synthesis pipeline to adapt existing large-scale 3D models for 3D anomaly detection. Specifically we construct a synthetic dataset i.e. Anomaly-ShapeNet based on ShapeNet. Anomaly-ShapeNet consists of 1600 point cloud samples under 40 categories which provides a rich and varied

collection of data enabling efficient training and enhancing adaptability to ind
ustrial scenarios. Meanwhile to enable scalable representation learning for 3D a
nomaly localization we propose a self-supervised method i.e. Iterative Mask Reco
nstruction Network (IMRNet). During training we propose a geometry-aware sample
module to preserve potentially anomalous local regions during point cloud down-s
ampling. Then we randomly mask out point patches and sent the visible patches to
 a transformer for reconstruction-based self-supervision. During testing the poi
nt cloud repeatedly goes through the Mask Reconstruction Network with each itera
tion's output becoming the next input. By merging and contrasting the final reco
nstructed point cloud with the initial input our method successfully locates ano
malies. Experiments show that IMRNet outperforms previous state-of-the-art metho
ds achieving 66.1% in I-AUC on our Anomaly-ShapeNet dataset and 72.5% in I-AUC o
n Real3D-AD dataset. Our benchmark will be released at https://github.com/Choppe
r233/Anomaly-ShapeNet.
****
Cam4DOcc: Benchmark for Camera-Only 4D Occupancy Forecasting in Autonomous Drivi
ng Applications
Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao G
u, Rui Ai, Hesheng Wang; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 21486-21495
Understanding how the surrounding environment changes is crucial for performing
downstream tasks safely and reliably in autonomous driving applications. Recent
occupancy estimation techniques using only camera images as input can provide de
nse occupancy representations of large-scale scenes based on the current observa
tion. However they are mostly limited to representing the current 3D space and d
o not consider the future state of surrounding objects along the time axis. To e
xtend camera-only occupancy estimation into spatiotemporal prediction we propose
 Cam4DOcc a new benchmark for camera-only 4D occupancy forecasting evaluating th
e surrounding scene changes in a near future. We build our benchmark based on mu
ltiple publicly available datasets including nuScenes nuScenes-Occupancy and Lyf
t-Level5 which provides sequential occupancy states of general movable and stati
c objects as well as their 3D backward centripetal flow. To establish this bench
mark for future research with comprehensive comparisons we introduce four baseli
ne types from diverse camera-based perception and prediction implementations inc
luding a static-world occupancy model voxelization of point cloud prediction 2D-
3D instance-based prediction and our proposed novel end-to-end 4D occupancy fore
casting network. Furthermore the standardized evaluation protocol for preset mul
tiple tasks is also provided to compare the performance of all the proposed base
lines on present and future occupancy estimation with respect to objects of inte
rest in autonomous driving scenarios. The dataset and our implementation of all
four baselines in the proposed Cam4DOcc benchmark are released as open source at
 https://github.com/haomo-ai/Cam4DOcc.
****
DIOD: Self-Distillation Meets Object Discovery
Sandra Kara, Hejer Ammar, Julien Denize, Florian Chabot, Quoc-Cuong Pham; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2024, pp. 3975-3985
Instance segmentation demands substantial labeling resources. This has prompted
increased interest to explore the object discovery task as an unsupervised alter
native. In particular promising results were achieved in localizing instances us
ing motion supervision only. However the motion signal introduces complexities d
ue to its inherent noise and sparsity which constrains the effectiveness of curr
ent methodologies. In the present paper we propose DIOD (self DIstillation meets
 Object Discovery) the first method that places the motion-guided object discove
ry within a framework of continuous improvement through knowledge distillation p
roviding solutions to existing limitations (i) DIOD robustly eliminates the nois
e present in the exploited motion maps providing accurate motion-supervision (ii
) DIOD leverages the discovered objects within an iterative pseudo-labeling fram
ework enriching the initial motion-supervision with static objects which results
 in a cost-efficient increase in performance. Through experiments on synthetic a

nd real-world datasets we demonstrate the benefits of bridging the gap between object discovery and distillation by significantly improving the state-of-the-art. This enhancement is also sustained across other demanding metrics so far reserved for supervised tasks.
****
GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh
Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G. Schwing, Shenlong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2059-2069
We introduce GoMAvatar a novel approach for real-time memory-efficient high-quality animatable human modeling. GoMAvatar takes as input a single monocular video to create a digital avatar capable of re-articulation in new poses and real-time rendering from novel viewpoints while seamlessly integrating with rasterization-based graphics pipelines. Central to our method is the Gaussians-on-Mesh (GoM) representation a hybrid 3D model combining rendering quality and speed of Gaussian splatting with geometry modeling and compatibility of deformable meshes. We assess GoMAvatar on ZJU-MoCap PeopleSnapshot and various YouTube videos. GoMAvatar matches or surpasses current monocular human modeling algorithms in rendering quality and significantly outperforms them in computational efficiency (43 FPS) while being memory-efficient (3.63 MB per subject).
****
Neural Redshift: Random Networks are not Random Functions
Damien Teney, Armand Mihai Nicolicioiu, Valentin Hartmann, Ehsan Abbasnejad; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4786-4796
Our understanding of the generalization capabilities of neural networks NNs is still incomplete. Prevailing explanations are based on implicit biases of gradient descent GD but they cannot account for the capabilities of models from gradientfree methods nor the simplicity bias recently observed in untrained networks This paper seeks other sources of generalization in NNs. To understand the inductive biases provided by architectures independently from GD we examine untrained randomweight networks Even simple MLPs show strong inductive biases uniform sampling in weight space yields a very biased distribution of functions in terms of complexity But unlike common wisdom NNs do not have an inherent simplicity bias This property depends on components such as ReLUs residual connections and layer normalizations Alternative architectures can be built with a bias for any level of complexity. Transformers also inherit all these properties from their building blocks. We provide a fresh explanation for the success of deep learning independent from gradientbased training It points at promising avenues for controlling the solutions implemented by trained models.
****
HumanGaussian: Text-Driven 3D Human Generation with Gaussian Splatting
Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6646-6657
Realistic 3D human generation from text prompts is a desirable yet challenging task. Existing methods optimize 3D representations like mesh or neural fields via score distillation sampling (SDS) which suffers from inadequate fine details or excessive training time. In this paper we propose an efficient yet effective framework HumanGaussian that generates high-quality 3D humans with fine-grained geometry and realistic appearance. Our key insight is that 3D Gaussian Splatting is an efficient renderer with periodic Gaussian shrinkage or growing where such adaptive density control can be naturally guided by intrinsic human structures. Specifically 1) we first propose a Structure-Aware SDS that simultaneously optimizes human appearance and geometry. The multi-modal score function from both RGB and depth space is leveraged to distill the Gaussian densification and pruning process. 2) Moreover we devise an Annealed Negative Prompt Guidance by decomposing SDS into a noisier generative score and a cleaner classifier score which well addresses the over-saturation issue. The floating artifacts are further eliminat

ed based on Gaussian size in a prune-only phase to enhance generation smoothness. Extensive experiments demonstrate the superior efficiency and competitive quality of our framework rendering vivid 3D humans under diverse scenarios.
****
DIEM: Decomposition-Integration Enhancing Multimodal Insights
Xinyi Jiang, Guoming Wang, Junhao Guo, Juncheng Li, Wenqiao Zhang, Rongxing Lu, Siliang Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27304-27313
In image question answering due to the abundant and sometimes redundant information precisely matching and integrating the information from both text and images is a challenge. In this paper we propose the Decomposition-Integration Enhancing Multimodal Insight (DIEM) which initially decomposes the given question and image into multiple subquestions and several sub-images aiming to isolate specific elements for more focused analysis. We then integrate these sub-elements by matching each subquestion with its relevant sub-images while also retaining the original image to construct a comprehensive answer to the original question without losing sight of the overall context. This strategy mirrors the human cognitive process of simplifying complex problems into smaller components for individual analysis followed by an integration of these insights. We implement DIEM on the LLaVA-v1.5 model and evaluate its performance on ScienceQA and MM-Vet. Experimental results indicate that our method boosts accuracy in most question classes of the ScienceQA (+2.03% in average) especially in the image modality (+3.40%). On MM-Vet our method achieves an improvement in MM-Vet scores increasing from 31.1 to 32.4. These findings highlight DIEM's effectiveness in harmonizing the complexities of multimodal data demonstrating its ability to enhance accuracy and depth in image question answering through its decomposition-integration process.
****
CosmicMan: A Text-to-Image Foundation Model for Humans
Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, Wayne Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6955-6965
We present CosmicMan a text-to-image foundation model specialized for generating high-fidelity human images. Unlike current general-purpose foundation models that are stuck in the dilemma of inferior quality and text-image misalignment for humans CosmicMan enables generating photo-realistic human images with meticulous appearance reasonable structure and precise text-image alignment with detailed dense descriptions. At the heart of CosmicMan's success are the new reflections and perspectives on data and models: (1) We found that data quality and a scalable data production flow are essential for the final results from trained models. Hence we propose a new data production paradigm Annotate Anyone which serves as a perpetual data flywheel to produce high-quality data with accurate yet cost-effective annotations over time. Based on this we constructed a large-scale dataset CosmicMan-HQ 1.0 with 6 Million high-quality real-world human images in a mean resolution of 1488x1255 and attached with precise text annotations deriving from 115 Million attributes in diverse granularities. (2) We argue that a text-to-image foundation model specialized for humans must be pragmatic - easy to integrate into down-streaming tasks while effective in producing high-quality human images. Hence we propose to model the relationship between dense text descriptions and image pixels in a decomposed manner and present Decomposed-Attention-Refocusing (Daring) training framework. It seamlessly decomposes the cross-attention features in existing text-to-image diffusion model and enforces attention refocusing without adding extra modules. Through Daring we show that explicitly discretizing continuous text space into several basic groups that align with human body structure is the key to tackling the misalignment problem in a breeze. Project page: https://cosmicman-cvpr2024.github.io/.
****
LLMs are Good Sign Language Translators
Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18362-18372

Sign Language Translation (SLT) is a challenging task that aims to translate sign videos into spoken language. Inspired by the strong translation capabilities of large language models (LLMs) that are trained on extensive multilingual text corpora we aim to harness off-the-shelf LLMs to handle SLT. In this paper we regularize the sign videos to embody linguistic characteristics of spoken language and propose a novel SignLLM framework to transform sign videos into a language-like representation for improved readability by off-the-shelf LLMs. SignLLM comprises two key modules: (1) The Vector-Quantized Visual Sign module converts sign videos into a sequence of discrete character-level sign tokens and (2) the Codebook Reconstruction and Alignment module converts these character-level tokens into word-level sign representations using an optimal transport formulation. A sign-text alignment loss further bridges the gap between sign and text tokens enhancing semantic compatibility. We achieve state-of-the-art gloss-free results on two widely-used SLT benchmarks.
****
Contrastive Pre-Training with Multi-View Fusion for No-Reference Point Cloud Quality Assessment
Ziyu Shan, Yujie Zhang, Qi Yang, Haichen Yang, Yiling Xu, Jenq-Neng Hwang, Xiaozhong Xu, Shan Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25942-25951
No-reference point cloud quality assessment (NR-PCQA) aims to automatically evaluate the perceptual quality of distorted point clouds without available reference which have achieved tremendous improvements due to the utilization of deep neural networks. However learning-based NR-PCQA methods suffer from the scarcity of labeled data and usually perform suboptimally in terms of generalization. To solve the problem we propose a novel contrastive pre-training framework tailored for PCQA (CoPA) which enables the pre-trained model to learn quality-aware representations from unlabeled data. To obtain anchors in the representation space we project point clouds with different distortions into images and randomly mix their local patches to form mixed images with multiple distortions. Utilizing the generated anchors we constrain the pre-training process via a quality-aware contrastive loss following the philosophy that perceptual quality is closely related to both content and distortion. Furthermore in the model fine-tuning stage we propose a semantic-guided multi-view fusion module to effectively integrate the features of projected images from multiple perspectives. Extensive experiments show that our method outperforms the state-of-the-art PCQA methods on popular benchmarks. Further investigations demonstrate that CoPA can also benefit existing learning-based PCQA models.
****
JDEC: JPEG Decoding via Enhanced Continuous Cosine Coefficients
Woo Kyoung Han, Sunghoon Im, Jaedeok Kim, Kyong Hwan Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2784-2793
We propose a practical approach to JPEG image decoding utilizing a local implicit neural representation with continuous cosine formulation. The JPEG algorithm significantly quantizes discrete cosine transform (DCT) spectra to achieve a high compression rate inevitably resulting in quality degradation while encoding an image. We have designed a continuous cosine spectrum estimator to address the quality degradation issue that restores the distorted spectrum. By leveraging local DCT formulations our network has the privilege to exploit dequantization and upsampling simultaneously. Our proposed model enables decoding compressed images directly across different quality factors using a single pre-trained model without relying on a conventional JPEG decoder. As a result our proposed network achieves state-of-the-art performance in flexible color image JPEG artifact removal tasks. Our source code is available at https://github.com/WooKyoungHan/JDEC
****
Revisiting the Domain Shift and Sample Uncertainty in Multi-source Active Domain Transfer
Wenqiao Zhang, Zheqi Lv, Hao Zhou, Jia-Wei Liu, Juncheng Li, Mengze Li, Yunfei Li, Dongping Zhang, Yueting Zhuang, Siliang Tang; Proceedings of the IEEE/CVF Con

ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16751-16761
Active Domain Adaptation (ADA) aims to maximally boost model adaptation in a new target domain by actively selecting a limited number of target data to annotate. This setting neglects the more practical scenario where training data are collected from multiple sources. This motivates us to extend ADA from a single source domain to multiple source domains termed Multi-source Active Domain Adaptation (MADA). Not surprisingly we find that most traditional ADA methods cannot work directly in such a setting mainly due to the excessive domain gap introduced by all the source domains. Considering this we propose a Detective framework that comprehensively considers the domain shift between multi-source domains and target domains to detect the informative target samples. Specifically the Detective leverages a dynamic Domain Adaptation (DA) model that learns how to adapt the model's parameters to fit the union of multi-source domains. This enables an approximate single-source domain modeling by the dynamic model. We then comprehensively measure both domain uncertainty and predictive uncertainty in the target domain to detect informative target samples using evidential deep learning thereby mitigating uncertainty miscalibration. Experiments demonstrate that our solution outperforms existing methods by a considerable margin on three domain adaptation benchmarks.
****
Learning Continual Compatible Representation for Re-indexing Free Lifelong Person Re-identification
Zhenyu Cui, Jiahuan Zhou, Xun Wang, Manyu Zhu, Yuxin Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16614-16623
Lifelong Person Re-identification (L-ReID) aims to learn from sequentially collected data to match a person across different scenes. Once an L-ReID model is updated using new data all historical images in the gallery are required to be re-calculated to obtain new features for testing known as "re-indexing". However it is infeasible when raw images in the gallery are unavailable due to data privacy concerns resulting in incompatible retrieval between the query and the gallery features calculated by different models which causes significant performance degradation. In this paper we focus on a new task called Re-indexing Free Lifelong Person Re-identification (RFL-ReID) which requires achieving effective L-ReID without re-indexing raw images in the gallery. To this end we propose a Continual Compatible Representation (C2R) method which facilitates the query feature calculated by the continuously updated model to effectively retrieve the gallery feature calculated by the old model in a compatible manner. Specifically we design a Continual Compatible Transfer (CCT) network to continuously transfer and consolidate the old gallery feature into the new feature space. Besides a Balanced Compatible Distillation module is introduced to achieve compatibility by aligning the transferred feature space with the new feature space. Finally a Balanced Anti-forgetting Distillation module is proposed to eliminate the accumulated forgetting of old knowledge during the continual compatible transfer. Extensive experiments on several benchmark L-ReID datasets demonstrate the effectiveness of our method against state-of-the-art methods for both RFL-ReID and L-ReID tasks. The source code of this paper is available at https://github.com/PKU-ICST-MIPL/C2R_CVPR2024.
****
Revisiting Spatial-Frequency Information Integration from a Hierarchical Perspective for Panchromatic and Multi-Spectral Image Fusion
Jiangtong Tan, Jie Huang, Naishan Zheng, Man Zhou, Keyu Yan, Danfeng Hong, Feng Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25922-25931
Pan-sharpening is a super-resolution problem that essentially relies on spectra fusion of panchromatic (PAN) images and low-resolution multi-spectral (LRMS) images. The previous methods have validated the effectiveness of information fusion in the Fourier space of the whole image. However they haven't fully explored the Fourier relationships at different hierarchies between PAN and LRMS images. To this end we propose a Hierarchical Frequency Integration Network (HFIN) to faci

litate hierarchical Fourier information integration for pan-sharpening. Specifically our network consists of two designs: information stratification and information integration. For information stratification we hierarchically decompose PAN and LRMS information into spatial global Fourier and local Fourier information and fuse them independently. For information integration the above hierarchical fused information is processed to further enhance their relationships and undergo comprehensive integration. Our method extend a new space for exploring the relationships of PAN and LRMS images enhancing the integration of spatial-frequency information. Extensive experiments robustly validate the effectiveness of the proposed network showcasing its superior performance compared to other state-of-the-art methods and generalization in real-world scenes and other fusion tasks as a general image fusion framework. Code is available at https://github.com/JosephTiTan/HFIN.
****
BSNet: Box-Supervised Simulation-assisted Mean Teacher for 3D Instance Segmentation
Jiahao Lu, Jiacheng Deng, Tianzhu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20374-20384
3D instance segmentation (3DIS) is a crucial task but point-level annotations are tedious in fully supervised settings. Thus using bounding boxes (bboxes) as annotations has shown great potential. The current mainstream approach is a two-step process involving the generation of pseudo-labels from box annotations and the training of a 3DIS network with the pseudo-labels. However due to the presence of intersections among bboxes not every point has a determined instance label especially in overlapping areas. To generate higher quality pseudo-labels and achieve more precise weakly supervised 3DIS results we propose the Box-Supervised Simulation-assisted Mean Teacher for 3D Instance Segmentation (BSNet) which devises a novel pseudo-labeler called Simulation-assisted Transformer. The labeler consists of two main components. The first is Simulation-assisted Mean Teacher which introduces Mean Teacher for the first time in this task and constructs simulated samples to assist the labeler in acquiring prior knowledge about overlapping areas. To better model local-global structure we also propose Local-Global Aware Attention as the decoder for teacher and student labelers. Extensive experiments conducted on the ScanNetV2 and S3DIS datasets verify the superiority of our designs.
****
Adaptive Slot Attention: Object Discovery with Dynamic Slot Number
Ke Fan, Zechen Bai, Tianjun Xiao, Tong He, Max Horn, Yanwei Fu, Francesco Locatello, Zheng Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23062-23071
Object-centric learning (OCL) extracts the representation of objects with slots offering an exceptional blend of flexibility and interpretability for abstracting low-level perceptual features. A widely adopted method within OCL is slot attention which utilizes attention mechanisms to iteratively refine slot representations. However a major drawback of most object-centric models including slot attention is their reliance on predefining the number of slots. This not only necessitates prior knowledge of the dataset but also overlooks the inherent variability in the number of objects present in each instance. To overcome this fundamental limitation we present a novel complexity-aware object auto-encoder framework. Within this framework we introduce an adaptive slot attention (AdaSlot) mechanism that dynamically determines the optimal number of slots based on the content of the data. This is achieved by proposing a discrete slot sampling module that is responsible for selecting an appropriate number of slots from a candidate list. Furthermore we introduce a masked slot decoder that suppresses unselected slots during the decoding process. Our framework tested extensively on object discovery tasks with various datasets shows performance matching or exceeding top fixed-slot models. Moreover our analysis substantiates that our method exhibits the capability to dynamically adapt the slot number according to each instance's complexity offering the potential for further exploration in slot attention research. Project will be available at https://kfan21.github.io/AdaSlot/

****
CORES: Convolutional Response-based Score for Out-of-distribution Detection
Keke Tang, Chao Hou, Weilong Peng, Runnan Chen, Peican Zhu, Wenping Wang, Zhihon
g Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2024, pp. 10916-10925
Deep neural networks (DNNs) often display overconfidence when encountering out-o
f-distribution (OOD) samples posing significant challenges in real-world applica
tions. Capitalizing on the observation that responses on convolutional kernels a
re generally more pronounced for in-distribution (ID) samples than for OOD ones
this paper proposes the COnvolutional REsponse-based Score (CORES) to exploit th
ese discrepancies for OOD detection. Initially CORES delves into the extremities
 of convolutional responses by considering both their magnitude and the frequenc
y of significant values. Moreover through backtracking from the most prominent p
redictions CORES effectively pinpoints sample-relevant kernels across different
layers. These kernels which exhibit a strong correlation to input samples are in
tegral to CORES's OOD detection capability. Comprehensive experiments across var
ious ID and OOD settings demonstrate CORES's effectiveness in OOD detection and
its superiority to the state-of-the-art methods.
****
Task-Driven Wavelets using Constrained Empirical Risk Minimization
Eric Marcus, Ray Sheombarsing, Jan-Jakob Sonke, Jonas Teuwen; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 24098-24107
Deep Neural Networks (DNNs) are widely used for their ability to effectively app
roximate large classes of functions. This flexibility however makes the strict e
nforcement of constraints on DNNs a difficult problem. In contexts where it is c
ritical to limit the function space to which certain network components belong s
uch as wavelets employed in Multi-Resolution Analysis (MRA) naive constraints vi
a additional terms in the loss function are inadequate. To address this we intro
duce a Convolutional Neural Network (CNN) wherein the convolutional filters are
strictly constrained to be wavelets. This allows the filters to update to task-o
ptimized wavelets during the training procedure. Our primary contribution lies i
n the rigorous formulation of these filters via a constrained empirical risk min
imization framework thereby providing an exact mechanism to enforce these struct
ural constraints. While our work is grounded in theory we investigate our approa
ch empirically through applications in medical imaging particularly in the task
of contour prediction around various organs achieving superior performance compa
red to baseline methods.
****
HOI-M^3: Capture Multiple Humans and Objects Interaction within Contextual Envir
onment
Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jing
yi Yu, Lan Xu, Jingya Wang; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2024, pp. 516-526
Humans naturally interact with both others and the surrounding multiple objects
engaging in various social activities. However recent advances in modeling human
-object interactions mostly focus on perceiving isolated individuals and objects
 due to fundamental data scarcity. In this paper we introduce HOI-M^3 a novel la
rge-scale dataset for modeling the interactions of Multiple huMans and Multiple
objects. Notably it provides accurate 3D tracking for both humans and objects fr
om dense RGB and object-mounted IMU inputs covering 199 sequences and 181M frame
s of diverse humans and objects under rich activities. With the unique HOI-M^3 d
ataset we introduce two novel data-driven tasks with companion strong baselines:
 monocular capture and unstructured generation of multiple human-object interact
ions. Extensive experiments demonstrate that our dataset is challenging and wort
hy of further research about multiple human-object interactions and behavior ana
lysis. Our HOI-M^3 dataset corresponding codes and pre-trained models will be di
sseminated to the community for future research.
****
Interactive3D: Create What You Want by Interactive 3D Generation

Shaocong Dong, Lihe Ding, Zhanpeng Huang, Zibin Wang, Tianfan Xue, Dan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4999-5008
3D object generation has undergone significant advancements yielding high-quality results. However fall short in achieving precise user control often yielding results that do not align with user expectations thus limiting their applicability. User-envisioning 3D object generation faces significant challenges in realizing its concepts using current generative models due to limited interaction capabilities. Existing methods mainly offer two approaches: (i) interpreting textual instructions with constrained controllability or (ii) reconstructing 3D objects from 2D images. Both of them limit customization to the confines of the 2D reference and potentially introduce undesirable artifacts during the 3D lifting process restricting the scope for direct and versatile 3D modifications. In this work we introduce Interactive3D an innovative framework for interactive 3D generation that grants users precise control over the generative process through extensive 3D interaction capabilities. Interactive3D is constructed in two cascading stages utilizing distinct 3D representations. The first stage employs Gaussian Splatting for direct user interaction allowing modifications and guidance of the generative direction at any intermediate step through (i) Adding and Removing components (ii) Deformable and Rigid Dragging (iii) Geometric Transformations and (iv) Semantic Editing. Subsequently the Gaussian splats are transformed into InstantNGP. We introduce a novel (v) Interactive Hash Refinement module to further add details and extract the geometry in the second stage. Our experiments demonstrate that proposed Interactive3D markedly improves the controllability and quality of 3D generation. Our project webpage is available at https://interactive-3d.github.io/.
****
DeiT-LT: Distillation Strikes Back for Vision Transformer Training on Long-Tailed Datasets
Harsh Rangwani, Pradipto Mondal, Mayank Mishra, Ashish Ramayee Asokan, R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23396-23406
Vision Transformer (ViT) has emerged as a prominent architecture for various computer vision tasks. In ViT we divide the input image into patch tokens and process them through a stack of self-attention blocks. However unlike Convolutional Neural Network (CNN) ViT's simple architecture has no informative inductive bias (e.g. locality etc.). Due to this ViT requires a large amount of data for pre-training. Various data-efficient approaches (DeiT) have been proposed to train ViT on balanced datasets effectively. However limited literature discusses the use of ViT for datasets with long-tailed imbalances. In this work we introduce DeiT-LT to tackle the problem of training ViTs from scratch on long-tailed datasets. In DeiT-LT we introduce an efficient and effective way of distillation from CNN via distillation \texttt DIST token by using out-of-distribution images and re-weighting the distillation loss to enhance focus on tail classes. This leads to the learning of local CNN-like features in early ViT blocks improving generalization for tail classes. Further to mitigate overfitting we propose distilling from a flat CNN teacher which leads to learning low-rank generalizable features for DIST tokens across all ViT blocks. With the proposed DeiT-LT scheme the distillation DIST token becomes an expert on the tail classes and the classifier CLS token becomes an expert on the head classes. The experts help to effectively learn features corresponding to both the majority and minority classes using a distinct set of tokens within the same ViT architecture. We show the effectiveness of DeiT-LT for training ViT from scratch on datasets ranging from small-scale CIFAR-10 LT to large-scale iNaturalist-2018. Project Page: https://rangwani-harsh.github.io/DeiT-LT.
****
Accurate Spatial Gene Expression Prediction by Integrating Multi-Resolution Features
Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, Joo Sang Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 115

91-11600
Recent advancements in Spatial Transcriptomics (ST) technology have facilitated detailed gene expression analysis within tissue contexts. However the high costs and methodological limitations of ST necessitate a more robust predictive model. In response this paper introduces TRIPLEX a novel deep learning framework designed to predict spatial gene expression from Whole Slide Images (WSIs). TRIPLEX uniquely harnesses multi-resolution features capturing cellular morphology at individual spots the local context around these spots and the global tissue organization. By integrating these features through an effective fusion strategy TRIPLEX achieves accurate gene expression prediction. Our comprehensive benchmark study conducted on three public ST datasets and supplemented with Visium data from 10X Genomics demonstrates that TRIPLEX outperforms current state-of-the-art models in Mean Squared Error (MSE) Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC). The model's predictions align closely with ground truth gene expression profiles and tumor annotations underscoring TRIPLEX's potential in advancing cancer diagnosis and treatment.
****
FCS: Feature Calibration and Separation for Non-Exemplar Class Incremental Learning
Qiwei Li, Yuxin Peng, Jiahuan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28495-28504
Non-Exemplar Class Incremental Learning (NECIL) involves learning a classification model on a sequence of data without access to exemplars from previously encountered old classes. Such a stringent constraint always leads to catastrophic forgetting of the learned knowledge. Currently existing methods either employ knowledge distillation techniques or preserved class prototypes to sustain prior knowledge. However two critical issues still persist. On the one hand as the model is continually updated the preserved prototypes of old classes will inevitably derive from the suitable location in the feature space of the new model. On the other hand due to the lack of exemplars the features of new classes will take the place of similar old classes which breaks the classification boundary. To address these challenges we propose a Feature Calibration and Separation (FCS) method for NECIL. Our approach comprises a Feature Calibration Network (FCN) that adapts prototypes of old classes to the new model via optimal transport learning approximating the drift of prototypes caused by model evolution. Additionally we also propose a Prototype-Involved Contrastive Loss (PIC) that enhances feature separation among different classes. Specifically to mitigate the boundary distortion arising from the interplay of classes from different learning stages prototypes are involved in pushing the feature of new classes away from the old classes. Extensive experiments on three datasets with different settings have demonstrated the superiority of our FCS method against the state-of-the-art class incremental learning approaches. Code is available at https://github.com/zhoujiahuan1991/CVPR2024-FCS.
****
Task2Box: Box Embeddings for Modeling Asymmetric Task Relationships
Rangel Daroya, Aaron Sun, Subhransu Maji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28827-28837
Modeling and visualizing relationships between tasks or datasets is an important step towards solving various meta-tasks such as dataset discovery multi-tasking and transfer learning. However many relationships such as containment and transferability are naturally asymmetric and current approaches for representation and visualization (e.g. t-SNE) do not readily support this. We propose Task2Box an approach to represent tasks using box embeddings---axis-aligned hyperrectangles in low dimensional spaces---that can capture asymmetric relationships between them through volumetric overlaps. We show that Task2Box accurately predicts unseen hierarchical relationships between nodes in ImageNet and iNaturalist datasets as well as transferability between tasks in the Taskonomy benchmark. We also show that box embeddings estimated from task representations (e.g. CLIP Task2Vec or attribute based) can be used to predict relationships between unseen tasks more accurately than classifiers trained on the same representations as well as hand

crafted asymmetric distances (e.g. KL divergence). This suggests that low-dimens
ional box embeddings can effectively capture these task relationships and have t
he added advantage of being interpretable. We use the approach to visualize rela
tionships among publicly available image classification datasets on popular data
set hosting platform called Hugging Face.
****
Behind the Veil: Enhanced Indoor 3D Scene Reconstruction with Occluded Surfaces
Completion
Su Sun, Cheng Zhao, Yuliang Guo, Ruoyu Wang, Xinyu Huang, Yingjie Victor Chen, L
iu Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2024, pp. 12744-12753
In this paper we present a novel indoor 3D reconstruction method with occluded s
urface completion given a sequence of depth readings. Prior state-of-the-art (SO
TA) methods only focus on the reconstruction of the visible areas in a scene neg
lecting the invisible areas due to the occlusions e.g. the contact surface betwe
en furniture occluded wall and floor. Our method tackles the task of completing
the occluded scene surfaces resulting in a complete 3D scene mesh. The core idea
 of our method is learning 3D geometry prior from various complete scenes to inf
er the occluded geometry of an unseen scene from solely depth measurements. We d
esign a coarse-fine hierarchical octree representation coupled with a dual-decod
er architecture i.e. Geo-decoder and 3D Inpainter which jointly reconstructs the
 complete 3D scene geometry. The Geo-decoder with detailed representation at fin
e levels is optimized online for each scene to reconstruct visible surfaces. The
 3D Inpainter with abstract representation at coarse levels is trained offline u
sing various scenes to complete occluded surfaces. As a result while the Geo-dec
oder is specialized for an individual scene the 3D Inpainter can be generally ap
plied across different scenes. We evaluate the proposed method on the 3D Complet
ed Room Scene (3D-CRS) and iTHOR datasets significantly outperforming the SOTA m
ethods by a gain of 16.8% and 24.2% in terms of the completeness of 3D reconstru
ction. 3D-CRS dataset including a complete 3D mesh of each scene is provided at
project webpage.
****
VideoGrounding-DINO: Towards Open-Vocabulary Spatio-Temporal Video Grounding
Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, Fahad Shahbaz K
han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2024, pp. 18909-18918
Video grounding aims to localize a spatio-temporal section in a video correspond
ing to an input text query. This paper addresses a critical limitation in curren
t video grounding methodologies by introducing an Open-Vocabulary Spatio-Tempora
l Video Grounding task. Unlike prevalent closed-set approaches that struggle wit
h open-vocabulary scenarios due to limited training data and predefined vocabula
ries our model leverages pre-trained representations from foundational spatial g
rounding models. This empowers it to effectively bridge the semantic gap between
 natural language and diverse visual content achieving strong performance in clo
sed-set and open-vocabulary settings. Our contributions include a novel spatio-t
emporal video grounding model surpassing state-of-the-art results in closed-set
evaluations on multiple datasets and demonstrating superior performance in open-
vocabulary scenarios. Notably the proposed model outperforms state-of-the-art me
thods in closed-set settings on VidSTG (Declarative and Interrogative) and HC-ST
VG (V1 and V2) datasets. Furthermore in open-vocabulary evaluations on HC-STVG V
1 and YouCook-Interactions our model surpasses the recent best-performing models
 by 4.88 m_vIoU and 1.83 accuracy demonstrating its efficacy in handling diverse
 linguistic and visual concepts for improved video understanding. Our codes will
 be publicly released.
****
OmniLocalRF: Omnidirectional Local Radiance Fields from Dynamic Videos
Dongyoung Choi, Hyeonjoong Jang, Min H. Kim; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6871-6880
Omnidirectional cameras are extensively used in various applications to provide
a wide field of vision. However they face a challenge in synthesizing novel view

s due to the inevitable presence of dynamic objects including the photographer i
n their wide field of view. In this paper we introduce a new approach called Omn
idirectional Local Radiance Fields (OmniLocalRF) that can render static-only sce
ne views removing and inpainting dynamic objects simultaneously. Our approach co
mbines the principles of local radiance fields with the bidirectional optimizati
on of omnidirectional rays. Our input is an omnidirectional video and we evaluat
e the mutual observations of the entire angle between the previous and current f
rames. To reduce ghosting artifacts of dynamic objects and inpaint occlusions we
 devise a multi-resolution motion mask prediction module. Unlike existing method
s that primarily separate dynamic components through the temporal domain our met
hod uses multi-resolution neural feature planes for precise segmentation which i
s more suitable for long 360-degree videos. Our experiments validate that OmniLo
calRF outperforms existing methods in both qualitative and quantitative metrics
especially in scenarios with complex real-world scenes. In particular our approa
ch eliminates the need for manual interaction such as drawing motion masks by ha
nd and additional pose estimation making it a highly effective and efficient sol
ution.
****
LoS: Local Structure-Guided Stereo Matching
Kunhong Li, Longguang Wang, Ye Zhang, Kaiwen Xue, Shunbo Zhou, Yulan Guo; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2024, pp. 19746-19756
Estimating disparities in challenging areas is difficult and limits the performa
nce of stereo matching models. In this paper we exploit local structure informat
ion (LSI) to enhance stereo matching. Specifically our LSI comprises a series of
 key elements including the slant plane (parameterised by disparity gradients) d
isparity offset details and neighbouring relations. This LSI empowers our method
 to effectively handle intricate structures including object boundaries and curv
ed surfaces. We bootstrap the LSI from monocular depth and subsequently iterativ
ely refine it to better capture the underlying scene geometry constraints. Build
ing upon the LSI we introduce the Local Structure-Guided Propagation (LSGP) whic
h enhances the disparity initialization optimization and refinement processes. B
y combining LSGP with a Gated Recurrent Unit (GRU) we present our novel stereo m
atching method referred to as Local Structure-guided stereo matching (LoS). Rema
rkably LoS achieves top-ranking results on four widely recognized public benchma
rk datasets (ETH3D Middlebury KITTI 15 & 12) demonstrating the superior capabili
ties of our proposed model.
****
Semantic Human Mesh Reconstruction with Textures
Xiaoyu Zhan, Jianxin Yang, Yuanqi Li, Jie Guo, Yanwen Guo, Wenping Wang; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 142-152
The field of 3D detailed human mesh reconstruction has made significant progress
 in recent years. However current methods still face challenges when used in ind
ustrial applications due to unstable results low-quality meshes and a lack of UV
 unwrapping and skinning weights. In this paper we present SHERT a novel pipelin
e that can reconstruct semantic human meshes with textures and high-precision de
tails. SHERT applies semantic- and normal-based sampling between the detailed su
rface (e.g. mesh and SDF) and the corresponding SMPL-X model to obtain a partial
ly sampled semantic mesh and then generates the complete semantic mesh by our sp
ecifically designed self-supervised completion and refinement networks. Using th
e complete semantic mesh as a basis we employ a texture diffusion model to creat
e human textures that are driven by both images and texts. Our reconstructed mes
hes have stable UV unwrapping high-quality triangle meshes and consistent semant
ic information. The given SMPL-X model provides semantic information and shape p
riors allowing SHERT to perform well even with incorrect and incomplete inputs.
The semantic information also makes it easy to substitute and animate different
body parts such as the face body and hands. Quantitative and qualitative experim
ents demonstrate that SHERT is capable of producing high-fidelity and robust sem
antic meshes that outperform state-of-the-art methods.

****

Think Twice Before Selection: Federated Evidential Active Learning for Medical Image Analysis with Domain Shifts

Jiayi Chen, Benteng Ma, Hengfei Cui, Yong Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11439-11449

Federated learning facilitates the collaborative learning of a global model across multiple distributed medical institutions without centralizing data. Nevertheless the expensive cost of annotation on local clients remains an obstacle to effectively utilizing local data. To mitigate this issue federated active learning methods suggest leveraging local and global model predictions to select a relatively small amount of informative local data for annotation. However existing methods mainly focus on all local data sampled from the same domain making them unreliable in realistic medical scenarios with domain shifts among different clients. In this paper we make the first attempt to assess the informativeness of local data derived from diverse domains and propose a novel methodology termed Federated Evidential Active Learning (FEAL) to calibrate the data evaluation under domain shift. Specifically we introduce a Dirichlet prior distribution in both local and global models to treat the prediction as a distribution over the probability simplex and capture both aleatoric and epistemic uncertainties by using the Dirichlet-based evidential model. Then we employ the epistemic uncertainty to calibrate the aleatoric uncertainty. Afterward we design a diversity relaxation strategy to reduce data redundancy and maintain data diversity. Extensive experiments and analysis on five real multi-center medical image datasets demonstrate the superiority of FEAL over the state-of-the-art active learning methods in federated scenarios with domain shifts. The code will be available at https://github.com/JiayiChen815/FEAL.

****

Probing the 3D Awareness of Visual Foundation Models

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, Varun Jampani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21795-21806

Recent advances in large-scale pretraining have yielded visual foundation models with strong capabilities. Not only can recent models generalize to arbitrary images for their training task their intermediate representations are useful for other visual tasks such as detection and segmentation. Given that such models can classify delineate and localize objects in 2D we ask whether they also represent their 3D structure? In this work we analyze the 3D awareness of visual foundation models. We posit that 3D awareness implies that representations (1) encode the 3D structure of the scene and (2) consistently represent the surface across views. We conduct a series of experiments using task-specific probes and zero-shot inference procedures on frozen features. Our experiments reveal several limitations of the current models. Our code and analysis can be found at https://github.com/mbanani/probe3d.

****

PIA: Your Personalized Image Animator via Plug-and-Play Modules in Text-to-Image Models

Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, Kai Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7747-7756

Recent advancements in personalized text-to-image (T2I) models have revolutionized content creation empowering non-experts to generate stunning images with unique styles. While promising animating these personalized images with realistic motions poses significant challenges in preserving distinct styles high-fidelity details and achieving motion controllability by text. In this paper we present PIA a Personalized Image Animator that excels in aligning with condition images achieving motion controllability by text and the compatibility with various personalized T2I models without specific tuning. To achieve these goals PIA builds upon a base T2I model with well-trained temporal alignment layers allowing for the seamless transformation of any personalized T2I model into an image animation mo

del. A key component of PIA is the introduction of the condition module which ta
kes as inputs the condition frame and inter-frame affinity. This module leverage
s the affinity hint to transfer appearance information from the condition frame
to individual frames in the latent space. This design mitigates the challenges o
f appearance-related frame alignment within PIA and allows for a stronger focus
on aligning with motion-related guidance. To address the lack of a benchmark for
 this field we introduce AnimateBench a comprehensive benchmark comprising diver
se personalized T2I models curated images and motion-related prompts. We show ex
tensive evaluations and applications on AnimateBench to verify the superiority o
f PIA.
****
When Visual Grounding Meets Gigapixel-level Large-scale Scenes: Benchmark and Ap
proach
Tao Ma, Bing Bai, Haozhe Lin, Heyuan Wang, Yu Wang, Lin Luo, Lu Fang; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 22119-22128
Visual grounding refers to the process of associating natural language expressio
ns with corresponding regions within an image. Existing benchmarks for visual gr
ounding primarily operate within small-scale scenes with a few objects. Neverthe
less recent advances in imaging technology have enabled the acquisition of gigap
ixel-level images providing high-resolution details in large-scale scenes contai
ning numerous objects. To bridge this gap between imaging and computer vision be
nchmarks and make grounding more practically valuable we introduce a novel datas
et named GigaGrounding designed to challenge visual grounding models in gigapixe
l-level large-scale scenes. We extensively analyze and compare the dataset with
existing benchmarks demonstrating that GigaGrounding presents unique challenges
such as large-scale scene understanding gigapixel-level resolution significant v
ariations in object scales and the "multi-hop expressions". Furthermore we intro
duced a simple yet effective grounding approach which employs a "glance-to-zoom-
in" paradigm and exhibits enhanced capabilities for addressing the GigaGrounding
 task. The dataset is available at www.gigavision.ai.
****
NeRF Analogies: Example-Based Visual Attribute Transfer for NeRFs
Michael Fischer, Zhengqin Li, Thu Nguyen-Phuoc, Aljaz Bozic, Zhao Dong, Carl Mar
shall, Tobias Ritschel; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2024, pp. 4640-4650
A Neural Radiance Field (NeRF) encodes the specific relation of 3D geometry and
appearance of a scene. We here ask the question whether we can transfer the appe
arance from a source NeRF onto a target 3D geometry in a semantically meaningful
 way such that the resulting new NeRF retains the target geometry but has an app
earance that is an analogy to the source NeRF. To this end we generalize classic
 image analogies from 2D images to NeRFs. We leverage correspondence transfer al
ong semantic affinity that is driven by semantic features from large pre-trained
 2D image models to achieve multi-view consistent appearance transfer. Our metho
d allows exploring the mix-and-match product space of 3D geometry and appearance
. We show that our method outperforms traditional stylization-based methods and
that a large majority of users prefer our method over several typical baselines.
 Project page: https://mfischer-ucl.github.io/nerf_analogies
****
Mind Artist: Creating Artistic Snapshots with Human Thought
Jiaxuan Chen, Yu Qi, Yueming Wang, Gang Pan; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27207-27217
We introduce Mind Artist (MindArt) a novel and efficient neural decoding archite
cture to snap artistic photographs from our mind in a controllable manner. Recen
tly progress has been made in image reconstruction with non-invasive brain recor
dings but it's still difficult to generate realistic images with high semantic f
idelity due to the scarcity of data annotations. Unlike previous methods this wo
rk casts the neural decoding into optimal transport (OT) and representation deco
upling problems. Specifically under discrete OT theory we design a graph matchin
g-guided neural representation learning framework to seek the underlying corresp

ondences between conceptual semantics and neural signals which yields a natural and meaningful self-supervisory task. Moreover the proposed MindArt structured with multiple stand-alone modal branches enables the seamless incorporation of semantic representation into any visual style information thus leaving it to have multi-modal reconstruction and training-free semantic editing capabilities. By doing so the reconstructed images of MindArt have phenomenal realism both in terms of semantics and appearance. We compare our MindArt with leading alternatives and achieve SOTA performance in different decoding tasks. Importantly our approach can directly generate a series of stylized "mind snapshots" w/o extra optimizations which may open up more potential applications. Code is available at https://github.com/JxuanC/MindArt.
****

ViTamin: Designing Scalable Vision Models in the Vision-Language Era
Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, Liang-Chieh Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12954-12966
Recent breakthroughs in vision-language models (VLMs) start a new page in the vision community. The VLMs provide stronger and more generalizable feature embeddings compared to those from ImageNet-pretrained models thanks to the training on the large-scale Internet image-text pairs. However despite the amazing achievement from the VLMs vanilla Vision Transformers (ViTs) remain the default choice for the image encoder. Although pure transformer proves its effectiveness in the text encoding area it remains questionable whether it is also the case for image encoding especially considering that various types of networks are proposed on the ImageNet benchmark which unfortunately are rarely studied in VLMs. Due to small data/model scale the original conclusions of model design on ImageNet can be limited and biased. In this paper we aim at building an evaluation protocol of vision models in the vision-language era under the contrastive language-image pretraining (CLIP) framework. We provide a comprehensive way to benchmark different vision models covering their zero-shot performance and scalability in both model and training data sizes. To this end we introduce ViTamin a new vision models tailored for VLMs. ViTamin-L significantly outperforms ViT-L by 2.0% ImageNet zero-shot accuracy when using the same publicly available DataComp-1B dataset and the same OpenCLIP training scheme. ViTamin-L presents promising results on 60 diverse benchmarks including classification retrieval open-vocabulary detection and segmentation and large multi-modal models. When further scaling up the model size our ViTamin-XL with only 436M parameters attains 82.9% ImageNet zero-shot accuracy surpassing 82.0% achieved by EVA-E that has ten times more parameters (4.4B).
****

Accept the Modality Gap: An Exploration in the Hyperbolic Space
Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, Ajanthan Thalaiyasingam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27263-27272
Recent advancements in machine learning have spotlighted the potential of hyperbolic spaces as they effectively learn hierarchical feature representations. While there has been progress in leveraging hyperbolic spaces in single-modality contexts its exploration in multimodal settings remains under explored. Some recent efforts have sought to transpose Euclidean multimodal learning techniques to hyperbolic spaces by adopting geodesic distance based contrastive losses. However we show both theoretically and empirically that such spatial proximity based contrastive loss significantly disrupts hierarchies in the latent space. To remedy this we advocate that the cross-modal representations should accept the inherent modality gap between text and images and introduce a novel approach to measure cross-modal similarity that does not enforce spatial proximity. Our approach show remarkable capabilities in preserving unimodal hierarchies while aligning the two modalities. Our experiments on a series of downstream tasks demonstrate that better latent structure emerges with our objective function while being superior in text-to-image and image-to-text retrieval tasks.
****

Unraveling Instance Associations: A Closer Look for Audio-Visual Segmentation
Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, Gustavo Carneiro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26497-26507
Audio-visual segmentation (AVS) is a challenging task that involves accurately segmenting sounding objects based on audio-visual cues. The effectiveness of audio-visual learning critically depends on achieving accurate cross-modal alignment between sound and visual objects. Successful audio-visual learning requires two essential components: 1) a challenging dataset with high-quality pixel-level multi-class annotated images associated with audio files and 2) a model that can establish strong links between audio information and its corresponding visual object. However these requirements are only partially addressed by current methods with training sets containing biased audio-visual data and models that generalise poorly beyond this biased training set. In this work we propose a new cost-effective strategy to build challenging and relatively unbiased high-quality audio-visual segmentation benchmarks. We also propose a new informative sample mining method for audio-visual supervised contrastive learning to leverage discriminative contrastive samples to enforce cross-modal understanding. We show empirical results that demonstrate the effectiveness of our benchmark. Furthermore experiments conducted on existing AVS datasets and on our new benchmark show that our method achieves state-of-the-art (SOTA) segmentation accuracy.
****
Few-Shot Object Detection with Foundation Models
Guangxing Han, Ser-Nam Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28608-28618
Few-shot object detection (FSOD) aims to detect objects with only a few training examples. Visual feature extraction and query-support similarity learning are the two critical components. Existing works are usually developed based on ImageNet pre-trained vision backbones and design sophisticated metric-learning networks for few-shot learning but still have inferior accuracy. In this work we study few-shot object detection using modern foundation models. First vision-only contrastive pre-trained DINOv2 model is used for the vision backbone which shows strong transferable performance without tuning the parameters. Second Large Language Model (LLM) is employed for contextualized few-shot learning with the input of all classes and query image proposals. Language instructions are carefully designed to prompt the LLM to classify each proposal in context. The contextual information include proposal-proposal relations proposal-class relations and class-class relations which can largely promote few-shot learning. We comprehensively evaluate the proposed model (FM-FSOD) in multiple FSOD benchmarks achieving state-of-the-arts performance.
****
FedMef: Towards Memory-efficient Federated Dynamic Pruning
Hong Huang, Weiming Zhuang, Chen Chen, Lingjuan Lyu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27548-27557
Federated learning (FL) promotes decentralized training while prioritizing data confidentiality. However its application on resource-constrained devices is challenging due to the high demand for computation and memory resources to train deep learning models. Neural network pruning techniques such as dynamic pruning could enhance model efficiency but directly adopting them in FL still poses substantial challenges including post-pruning performance degradation high activation memory usage etc. To address these challenges we propose FedMef a novel and memory-efficient federated dynamic pruning framework. FedMef comprises two key components. First we introduce the budget-aware extrusion that maintains pruning efficiency while preserving post-pruning performance by salvaging crucial information from parameters marked for pruning within a given budget. Second we propose scaled activation pruning to effectively reduce activation memory footprints which is particularly beneficial for deploying FL to memory-limited devices. Extensive experiments demonstrate the effectiveness of our proposed FedMef. In particular it achieves a significant reduction of 28.5% in memory footprint compared to st

ate-of-the-art methods while obtaining superior accuracy.
****
Seeing the Unseen: Visual Common Sense for Semantic Placement
Ram Ramrakhya, Aniruddha Kembhavi, Dhruv Batra, Zsolt Kira, Kuo-Hao Zeng, Luca Weihs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16273-16283
Computer vision tasks typically involve describing what is visible in an image (e.g. classification detection segmentation and captioning). We study a visual common sense task that requires understanding 'what is not visible'. Specifically given an image (e.g. of a living room) and a name of an object ("cushion") a vision system is asked to predict semantically-meaningful regions (masks or bounding boxes) in the image where that object could be placed or is likely be placed by humans (e.g. on the sofa). We call this task: Semantic Placement (SP) and believe that such common-sense visual understanding is critical for assitive robots (tidying a house) AR devices (automatically rendering an object in the user's space) and visually-grounded chatbots with common sense. Studying the invisible is hard. Datasets for image description are typically constructed by curating relevant images (e.g. via image search with object names) and asking humans to annotate the contents of the image; neither of those two steps are straightforward for objects not present in the image. We overcome this challenge by operating in the opposite direction: we start with an image of an object in context (which is easy to find online) and remove that object from the image via inpainting. This automated pipeline converts unstructured web data into a paired with/without object dataset. With this proposed data generation pipeline we collect a novel dataset containing 1.3M images across 9 object categories. We then train a SP prediction model called CLIP-UNet on our dataset. The CLIP-UNet outperforms existing VLMs and baselines that combine semantic priors with object detectors generalizes well to real-world and simulated images exhibits semantics-aware reasoning for object placement and enables downstream applications like tidying robots in indoor environments.
****
Texture-Preserving Diffusion Models for High-Fidelity Virtual Try-On
Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, Xiangmin Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7017-7026
Image-based virtual try-on is an increasingly important task for online shopping. It aims to synthesize images of a specific person wearing a specified garment. Diffusion model-based approaches have recently become popular as they are excellent at image synthesis tasks. However these approaches usually employ additional image encoders and rely on the cross-attention mechanism for texture transfer from the garment to the person image which affects the try-on's efficiency and fidelity. To address these issues we propose an Texture-Preserving Diffusion (TPD) model for virtual try-on which enhances the fidelity of the results and introduces no additional image encoders. Accordingly we make contributions from two aspects. First we propose to concatenate the masked person and reference garment images along the spatial dimension and utilize the resulting image as the input for the diffusion model's denoising UNet. This enables the original self-attention layers contained in the diffusion model to achieve efficient and accurate texture transfer. Second we propose a novel diffusion-based method that predicts a precise inpainting mask based on the person and reference garment images further enhancing the reliability of the try-on results. In addition we integrate mask prediction and image synthesis into a single compact model. The experimental results show that our approach can be applied to various try-on tasks e.g. garment-to-person and person-to-person try-ons and significantly outperforms state-of-the-art methods on popular VITON VITON-HD databases. Code is available at https://github.com/Gal4way/TPD.
****
PracticalDG: Perturbation Distillation on Vision-Language Models for Hybrid Domain Generalization
Zining Chen, Weiqiu Wang, Zhicheng Zhao, Fei Su, Aidong Men, Hongying Meng; Proc

eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 23501-23511
Domain Generalization (DG) aims to resolve distribution shifts between source an
d target domains and current DG methods are default to the setting that data fro
m source and target domains share identical categories. Nevertheless there exist
s unseen classes from target domains in practical scenarios. To address this iss
ue Open Set Domain Generalization (OSDG) has emerged and several methods have be
en exclusively proposed. However most existing methods adopt complex architectur
es with slight improvement compared with DG methods. Recently vision-language mo
dels (VLMs) have been introduced in DG following the fine-tuning paradigm but co
nsume huge training overhead with large vision models. Therefore in this paper w
e innovate to transfer knowledge from VLMs to lightweight vision models and impr
ove the robustness by introducing Perturbation Distillation (PD) from three pers
pectives including Score Class and Instance (SCI) named SCI-PD. Moreover previou
s methods are oriented by the benchmarks with identical and fixed splits ignorin
g the divergence between source domains. These methods are revealed to suffer fr
om sharp performance decay with our proposed new benchmark Hybrid Domain General
ization (HDG) and a novel metric H^ 2 -CV which construct various splits to comp
rehensively assess the robustness of algorithms. Extensive experiments demonstra
te that our method outperforms state-of-the-art algorithms on multiple datasets
especially improving the robustness when confronting data scarcity.
****
SODA: Bottleneck Diffusion Models for Representation Learning
Drew A. Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K. Lampinen, Andrew Jae
gle, James L. McClelland, Loic Matthey, Felix Hill, Alexander Lerchner; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2024, pp. 23115-23127
We introduce SODA a self-supervised diffusion model designed for representation
learning. The model incorporates an image encoder which distills a source view i
nto a compact representation that in turn guides the generation of related novel
 views. We show that by imposing a tight bottleneck between the encoder and a de
noising decoder and leveraging novel view synthesis as a self-supervised objecti
ve we can turn diffusion models into strong representation learners capable of c
apturing visual semantics in an unsupervised manner. To the best of our knowledg
e SODA is the first diffusion model to succeed at ImageNet linear-probe classifi
cation and at the same time it accomplishes reconstruction editing and synthesis
 tasks across a wide range of datasets. Further investigation reveals the disent
angled nature of its emergent latent space that serves as an effective interface
 to control and manipulate the produced images. All in all we aim to shed light
on the exciting and promising potential of diffusion models not only for image g
eneration but also for learning rich and robust representations. See our website
 at soda-diffusion.github.io.
****
Towards Robust Event-guided Low-Light Image Enhancement: A Large-Scale Real-Worl
d Event-Image Dataset and Novel Approach
Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, Lin Wang; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 23-33
Event camera has recently received much attention for low-light image enhancemen
t (LIE) thanks to their distinct advantages such as high dynamic range. However
current research is prohibitively restricted by the lack of large-scale real-wor
ld and spatial-temporally aligned event-image datasets. To this end we propose a
 real-world (indoor and outdoor) dataset comprising over 30K pairs of images and
 events under both low and normal illumination conditions. To achieve this we ut
ilize a robotic arm that traces a consistent non-linear trajectory to curate the
 dataset with spatial alignment precision under 0.03mm. We then introduce a matc
hing alignment strategy rendering 90% of our dataset with errors less than 0.01s
. Based on the dataset we propose a novel event-guided LIE approach called EvLig
ht towards robust performance in real-world low-light scenes. Specifically we fi
rst design the multi-scale holistic fusion branch to extract holistic structural

and textural information from both events and images. To ensure robustness agai
nst variations in the regional illumination and noise we then introduce a Signal
-to-Noise-Ratio (SNR)-guided regional feature selection to selectively fuse feat
ures of images from regions with high SNR and enhance those with low SNR by extr
acting regional structural information from events. our EvLight significantly su
rpasses the frame-based methods e.g. Retinexformer by 1.14 dB and 2.62 dB respec
tively. Code and datasets are available at https://vlislab22.github.io/eg-lowlig
ht/.
****
Zero-Reference Low-Light Enhancement via Physical Quadruple Priors
Wenjing Wang, Huan Yang, Jianlong Fu, Jiaying Liu; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26057-260
66
Understanding illumination and reducing the need for supervision pose a signific
ant challenge in low-light enhancement. Current approaches are highly sensitive
to data usage during training and illumination-specific hyper-parameters limitin
g their ability to handle unseen scenarios. In this paper we propose a new zero-
reference low-light enhancement framework trainable solely with normal light ima
ges. To accomplish this we devise an illumination-invariant prior inspired by th
e theory of physical light transfer. This prior serves as the bridge between nor
mal and low-light images. Then we develop a prior-to-image framework trained wit
hout low-light data. During testing this framework is able to restore our illumi
nation-invariant prior back to images automatically achieving low-light enhancem
ent. Within this framework we leverage a pretrained generative diffusion model f
or model ability introduce a bypass decoder to handle detail distortion as well
as offer a lightweight version for practicality. Extensive experiments demonstra
te our framework's superiority in various scenarios as well as good interpretabi
lity robustness and efficiency. Code is available on our project homepage: http:
//daooshee.github.io/QuadPrior-Website/
****
LLaMA-Excitor: General Instruction Tuning via Indirect Feature Interaction
Bo Zou, Chao Yang, Yu Qiao, Chengbin Quan, Youjian Zhao; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 140
89-14099
Existing methods to fine-tune LLMs like Adapter Prefix-tuning and LoRA which int
roduce extra modules or additional input sequences to inject new skills or knowl
edge may compromise the innate abilities of LLMs. In this paper we propose LLaMA
-Excitor a lightweight method that stimulates the LLMs' potential to better foll
ow instructions by gradually paying more attention to worthwhile information. Sp
ecifically the LLaMA-Excitor does not directly change the intermediate hidden st
ate during the self-attention calculation of the transformer structure. We desig
ned the Excitor block as a bypass module for the similarity score computation in
 LLMs' self-attention to reconstruct keys and change the importance of values by
 learnable prompts. LLaMA-Excitor ensures a self-adaptive allocation of addition
al attention to input instructions thus effectively preserving LLMs' pre-trained
 knowledge when fine-tuning LLMs on low-quality instruction-following datasets.
Furthermore we unify the modeling of multi-modal tuning and language-only tuning
 extending LLaMA-Excitor to a powerful visual instruction follower without the n
eed for complex multi-modal alignment. Our proposed approach is evaluated in lan
guage-only and multi-modal tuning experimental scenarios. Notably LLaMA-Excitor
is the only method that maintains basic capabilities while achieving a significa
nt improvement (+6%) on the MMLU benchmark. In the visual instruction tuning we
achieve a new state-of-the-art image captioning performance of 157.5 CIDEr on MS
COCO and a comparable performance (88.39%) on ScienceQA to cutting-edge models w
ith more parameters and extensive vision-language pertaining.
****
NeRFCodec: Neural Feature Compression Meets Neural Radiance Fields for Memory-Ef
ficient Scene Representation
Sicheng Li, Hao Li, Yiyi Liao, Lu Yu; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21274-21283

The emergence of Neural Radiance Fields (NeRF) has greatly impacted 3D scene mod
eling and novel-view synthesis. As a kind of visual media for 3D scene represent
ation compression with high rate-distortion performance is an eternal target. Mo
tivated by advances in neural compression and neural field representation we pro
pose NeRFCodec an end-to-end NeRF compression framework that integrates non-line
ar transform quantization and entropy coding for memory-efficient scene represen
tation. Since training a non-linear transform directly on a large scale of NeRF
feature planes is impractical we discover that pre-trained neural 2D image codec
 can be utilized for compressing the features when adding content-specific param
eters. Specifically we reuse neural 2D image codec but modify its encoder and de
coder heads while keeping the other parts of the pre-trained decoder frozen. Thi
s allows us to train the full pipeline via supervision of rendering loss and ent
ropy loss yielding the rate-distortion balance by updating the content-specific
parameters. At test time the bitstreams containing latent code feature decoder h
ead and other side information are transmitted for communication. Experimental r
esults demonstrate our method outperforms existing NeRF compression methods enab
ling high-quality novel view synthesis with a memory budget of 0.5 MB.
****
From a Bird's Eye View to See: Joint Camera and Subject Registration without the
 Camera Calibration
Zekun Qian, Ruize Han, Wei Feng, Song Wang; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 863-873
We tackle a new problem of multi-view camera and subject registration in the bir
d's eye view (BEV) without pre-given camera calibration which promotes the multi
-view subject registration problem to a new calibration-free stage. This greatly
 alleviates the limitation in many practical applications. However this is a ver
y challenging problem since its only input is several RGB images from different
first-person views (FPVs) without the BEV image and the calibration of the FPVs
while the output is a unified plane aggregated from all views with the positions
 and orientations of both the subjects and cameras in a BEV. For this purpose we
 propose an end-to-end framework solving camera and subject registration togethe
r by taking advantage of their mutual dependence whose main idea is as below: i)
 creating a subject view-transform module (VTM) to project each pedestrian from
FPV to a virtual BEV ii) deriving a multi-view geometry-based spatial alignment
module (SAM) to estimate the relative camera pose in a unified BEV iii) selectin
g and refining the subject and camera registration results within the unified BE
V. We collect a new large-scale synthetic dataset with rich annotations for trai
ning and evaluation. Additionally we also collect a real dataset for cross-domai
n evaluation. The experimental results show the remarkable effectiveness of our
method. The code and proposed datasets are available at https://github.com/zekun
qian/BEVSee.
****
Steerers: A Framework for Rotation Equivariant Keypoint Descriptors
Georg Bökman, Johan Edstedt, Michael Felsberg, Fredrik Kahl; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 4885-4895
Image keypoint descriptions that are discriminative and matchable over large cha
nges in viewpoint are vital for 3D reconstruction. However descriptions output b
y learned descriptors are typically not robust to camera rotation. While they ca
n be made more robust by e.g. data aug-mentation this degrades performance on up
right images. Another approach is test-time augmentation which incurs a signific
ant increase in runtime. Instead we learn a lin-ear transform in description spa
ce that encodes rotations of the input image. We call this linear transform a st
eerer since it allows us to transform the descriptions as if the im-age was rota
ted. From representation theory we know all possible steerers for the rotation g
roup. Steerers can be optimized (A) given a fixed descriptor (B) jointly with a
de-scriptor or (C) we can optimize a descriptor given a fixed steerer. We perfor
m experiments in these three settings and obtain state-of-the-art results on the
 rotation invariant im-age matching benchmarks AIMS and Roto-360.
****

Efficient Dataset Distillation via Minimax Diffusion
Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, Yiran Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15793-15803
Dataset distillation reduces the storage and computational consumption of training a network by generating a small surrogate dataset that encapsulates rich information of the original large-scale one. However previous distillation methods heavily rely on the sample-wise iterative optimization scheme. As the images-per-class (IPC) setting or image resolution grows larger the necessary computation will demand overwhelming time and resources. In this work we intend to incorporate generative diffusion techniques for computing the surrogate dataset. Observing that key factors for constructing an effective surrogate dataset are representativeness and diversity we design additional minimax criteria in the generative training to enhance these facets for the generated images of diffusion models. We present a theoretical model of the process as hierarchical diffusion control demonstrating the flexibility of the diffusion process to target these criteria without jeopardizing the faithfulness of the sample to the desired distribution. The proposed method achieves state-of-the-art validation performance while demanding much less computational resources. Under the 100-IPC setting on ImageWoof our method requires less than one-twentieth the distillation time of previous methods yet yields even better performance. Source code and generated data are available in https://github.com/vimar-gu/MinimaxDiffusion.
****
Posterior Distillation Sampling
Juil Koo, Chanho Park, Minhyuk Sung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13352-13361
We introduce Posterior Distillation Sampling (PDS) a novel optimization method for parametric image editing based on diffusion models. Existing optimization-based methods which leverage the powerful 2D prior of diffusion models to handle various parametric images have mainly focused on generation. Unlike generation editing requires a balance between conforming to the target attribute and preserving the identity of the source content. Recent 2D image editing methods have achieved this balance by leveraging the stochastic latent encoded in the generative process of diffusion models. To extend the editing capabilities of diffusion models shown in pixel space to parameter space we reformulate the 2D image editing method into an optimization form named PDS. PDS matches the stochastic latents of the source and the target enabling the sampling of targets in diverse parameter spaces that align with a desired attribute while maintaining the source's identity. We demonstrate that this optimization resembles running a generative process with the target attribute but aligning this process with the trajectory of the source's generative process. Extensive editing results in Neural Radiance Fields and Scalable Vector Graphics representations demonstrate that PDS is capable of sampling targets to fulfill the aforementioned balance across various parameter spaces.
****
HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields
Haozhe Qi, Chen Zhao, Mathieu Salzmann, Alexander Mathis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10392-10402
Human hands are highly articulated and versatile at handling objects. Jointly estimating the 3D poses of a hand and the object it manipulates from a monocular camera is challenging due to frequent occlusions. Thus existing methods often rely on intermediate 3D shape representations to increase performance. These representations are typically explicit such as 3D point clouds or meshes and thus provide information in the direct surroundings of the intermediate hand pose estimate. To address this we introduce HOISDF a Signed Distance Field (SDF) guided hand-object pose estimation network which jointly exploits hand and object SDFs to provide a global implicit representation over the complete reconstruction volume. Specifically the role of the SDFs is threefold: equip the visual encoder with i

mplicit shape information help to encode hand-object interactions and guide the hand and object pose regression via SDF-based sampling and by augmenting the fea ture representations. We show that HOISDF achieves state-of-the-art results on h and-object pose estimation benchmarks (DexYCB and HO3Dv2). Code is available htt ps://github.com/amathislab/HOISDF.
****
Enhancing Video Super-Resolution via Implicit Resampling-based Alignment
Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, Angela Yao; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2546-2555
In video super-resolution it is common to use a frame-wise alignment to support the propagation of information over time. The role of alignment is well-studied for low-level enhancement in video but existing works overlook a critical step -- resampling. We show through extensive experiments that for alignment to be eff ective the resampling should preserve the reference frequency spectrum while min imizing spatial distortions. However most existing works simply use a default ch oice of bilinear interpolation for resampling even though bilinear interpolation has a smoothing effect and hinders super-resolution. From these observations we propose an implicit resampling-based alignment. The sampling positions are enco ded by a sinusoidal positional encoding while the value is estimated with a coor dinate network and a window-based cross-attention. We show that bilinear interpo lation inherently attenuates high-frequency information while an MLP-based coord inate network can approximate more frequencies. Experiments on synthetic and rea l-world datasets show that alignment with our proposed implicit resampling enhan ces the performance of state-of-the-art frameworks with minimal impact on both c ompute and parameters.
****
DiffPortrait3D: Controllable Diffusion for Zero-Shot Portrait View Synthesis
Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, Li njie Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10456-10465
We present DiffPortrait3D a conditional diffusion model that is capable of synth esizing 3D-consistent photo-realistic novel views from as few as a single in-the -wild portrait. Specifically given a single RGB input we aim to synthesize plaus ible but consistent facial details rendered from novel camera views with retaine d both identity and facial expression. In lieu of time-consuming optimization an d fine-tuning our zero-shot method generalizes well to arbitrary face portraits with unposed camera views extreme facial expressions and diverse artistic depict ions. At its core we leverage the generative prior of 2D diffusion models pre-tr ained on large-scale image datasets as our rendering backbone while the denoisin g is guided with disentangled attentive control of appearance and camera pose. T o achieve this we first inject the appearance context from the reference image i nto the self-attention layers of the frozen UNets. The rendering view is then ma nipulated with a novel conditional control module that interprets the camera pos e by watching a condition image of a crossed subject from the same view. Further more we insert a trainable cross-view attention module to enhance view consisten cy which is further strengthened with a novel 3D-aware noise generation process during inference. We demonstrate state-of-the-art results both qualitatively and quantitatively on our challenging in-the-wild and multi-view benchmarks.
****
Rethinking Transformers Pre-training for Multi-Spectral Satellite Imagery
Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Kh an, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2024, pp. 27811-27819
Recent advances in unsupervised learning have demonstrated the ability of large vision models to achieve promising results on downstream tasks by pre-training o n large amount of unlabelled data. Such pre-training techniques have also been e xplored recently in the remote sensing domain due to the availability of large a mount of unlabelled data. Different from standard natural image datasets remote sensing data is acquired from various sensor technologies and exhibit diverse ra

nge of scale variations as well as modalities. Existing satellite image pre-training methods either ignore the scale information present in the remote sensing imagery or restrict themselves to use only a single type of data modality. In this paper we re-visit transformers pre-training and leverage multi-scale information that is effectively utilized with multiple modalities. Our proposed approach named SatMAE++ performs multi-scale pre-training and utilizes convolution based upsampling blocks to reconstruct the image at higher scales making it extensible to include more scales. Compared to existing works the proposed SatMAE++ with multi-scale pre-training is equally effective for both optical as well as multi-spectral imagery. Extensive experiments on six datasets reveal the merits of proposed contributions leading to state-of-the-art performance on all datasets. SatMAE++ achieves mean average precision (mAP) gain of 2.5% for multi-label classification task on BigEarthNet dataset.
****
LLM4SGG: Large Language Models for Weakly Supervised Scene Graph Generation
Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, Chanyoung Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28306-28316
Weakly-Supervised Scene Graph Generation (WSSGG) research has recently emerged as an alternative to the fully-supervised approach that heavily relies on costly annotations. In this regard studies on WSSGG have utilized image captions to obtain unlocalized triplets while primarily focusing on grounding the unlocalized triplets over image regions. However they have overlooked the two issues involved in the triplet formation process from the captions: 1) Semantic over-simplification issue arises when extracting triplets from captions where fine-grained predicates in captions are undesirably converted into coarse-grained predicates resulting in a long-tailed predicate distribution and 2) Low-density scene graph issue arises when aligning the triplets in the caption with entity/predicate classes of interest where many triplets are discarded and not used in training leading to insufficient supervision. To tackle the two issues we propose a new approach i.e. Large Language Model for weakly-supervised SGG (LLM4SGG) where we mitigate the two issues by leveraging the LLM's in-depth understanding of language and reasoning ability during the extraction of triplets from captions and alignment of entity/predicate classes with target data. To further engage the LLM in these processes we adopt the idea of Chain-of-Thought and the in-context few-shot learning strategy. To validate the effectiveness of LLM4SGG we conduct extensive experiments on Visual Genome and GQA datasets showing significant improvements in both Recall@K and mean Recall@K compared to the state-of-the-art WSSGG methods. A further appeal is that LLM4SGG is data-efficient enabling effective model training with a small amount of training images.
****
Parameter Efficient Fine-tuning via Cross Block Orchestration for Segment Anything Model
Zelin Peng, Zhengqin Xu, Zhilin Zeng, Lingxi Xie, Qi Tian, Wei Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3743-3752
Parameter-efficient fine-tuning (PEFT) is an effective methodology to unleash the potential of large foundation models in novel scenarios with limited training data. In the computer vision community PEFT has shown effectiveness in image classification but little research has studied its ability for image segmentation. Fine-tuning segmentation models usually require a heavier adjustment of parameters to align the proper projection directions in the parameter space for new scenarios. This raises a challenge to existing PEFT algorithms as they often inject a limited number of individual parameters into each block which prevents substantial adjustment of the projection direction of the parameter space due to the limitation of Hidden Markov Chain along blocks. In this paper we equip PEFT with a cross-block orchestration mechanism to enable the adaptation of the Segment Anything Model (SAM) to various downstream scenarios. We introduce a novel inter-block communication module which integrates a learnable relation matrix to facilitate communication among different coefficient sets of each PEFT block's paramete

r space. Moreover we propose an intra-block enhancement module which introduces a linear projection head whose weights are generated from a hyper-complex layer further enhancing the impact of the adjustment of projection directions on the entire parameter space. Extensive experiments on diverse benchmarks demonstrate that our proposed approach consistently improves the segmentation performance significantly on novel scenarios with only around 1K additional parameters.
****
Neural Directional Encoding for Efficient and Accurate View-Dependent Appearance Modeling
Liwen Wu, Sai Bi, Zexiang Xu, Fujun Luan, Kai Zhang, Iliyan Georgiev, Kalyan Sunkavalli, Ravi Ramamoorthi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21157-21166
Novel-view synthesis of specular objects like shiny metals or glossy paints remains a significant challenge. Not only the glossy appearance but also global illumination effects including reflections of other objects in the environment are critical components to faithfully reproduce a scene. In this paper we present Neural Directional Encoding (NDE) a view-dependent appearance encoding of neural radiance fields (NeRF) for rendering specular objects. NDE transfers the concept of feature-grid-based spatial encoding to the angular domain significantly improving the ability to model high-frequency angular signals. In contrast to previous methods that use encoding functions with only angular input we additionally cone-trace spatial features to obtain a spatially varying directional encoding which addresses the challenging interreflection effects. Extensive experiments on both synthetic and real datasets show that a NeRF model with NDE (1) outperforms the state of the art on view synthesis of specular objects and (2) works with small networks to allow fast (real-time) inference. The source code is available at: https://github.com/lwwu2/nde
****
Masked and Shuffled Blind Spot Denoising for Real-World Images
Hamadi Chihaoui, Paolo Favaro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3025-3034
We introduce a novel approach to single image denoising based on the Blind Spot Denoising principle which we call MAsked and SHuffled Blind Spot Denoising (MASH). We focus on the case of correlated noise which often plagues real images. MASH is the result of a careful analysis to determine the relationships between the level of blindness (masking) of the input and the (unknown) noise correlation. Moreover we introduce a shuffling technique to weaken the local correlation of noise which in turn yields an additional denoising performance improvement. We evaluate MASH via extensive experiments on real-world noisy image datasets. We demonstrate state-of-the-art results compared to existing self-supervised denoising methods.
****
Label Propagation for Zero-shot Classification with Vision-Language Models
Vladan Stojni?, Yannis Kalantidis, Giorgos Tolias; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23209-23218
Vision-Language Models (VLMs) have demonstrated impressive performance on zero-shot classification i.e. classification when provided merely with a list of class names. In this paper we tackle the case of zero-shot classification in the presence of unlabeled data. We leverage the graph structure of the unlabeled data and introduce ZLaP a method based on label propagation (LP) that utilizes geodesic distances for classification. We tailor LP to graphs containing both text and image features and further propose an efficient method for performing inductive inference based on a dual solution and a sparsification step. We perform extensive experiments to evaluate the effectiveness of our method on 14 common datasets and show that ZLaP outperforms the latest related works. Code: https://github.com/vladan-stojnic/ZLaP
****
DiffusionAvatars: Deferred Diffusion for High-fidelity 3D Head Avatars
Tobias Kirschstein, Simon Giebenhain, Matthias Nießner; Proceedings of the IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5481-5492
DiffusionAvatars synthesizes a high-fidelity 3D head avatar of a person offering intuitive control over both pose and expression. We propose a diffusion-based neural renderer that leverages generic 2D priors to produce compelling images of faces. For coarse guidance of the expression and head pose we render a neural parametric head model (NPHM) from the target viewpoint which acts as a proxy geometry of the person. Additionally to enhance the modeling of intricate facial expressions we condition DiffusionAvatars directly on the expression codes obtained from NPHM via cross-attention. Finally to synthesize consistent surface details across different viewpoints and expressions we rig learnable spatial features to the head's surface via TriPlane lookup in NPHM's canonical space. We train DiffusionAvatars on RGB videos and corresponding fitted NPHM meshes of a person and test the obtained avatars in both self-reenactment and animation scenarios. Our experiments demonstrate that DiffusionAvatars generates temporally consistent and visually appealing videos for novel poses and expressions of a person outperforming existing approaches.
****
Data-Free Quantization via Pseudo-label Filtering
Chunxiao Fan, Ziqi Wang, Dan Guo, Meng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5589-5598
Quantization for model compression can efficiently reduce the network complexity and storage requirement but the original training data is necessary to remedy the performance loss caused by quantization. The Data-Free Quantization (DFQ) methods have been proposed to handle the absence of original training data with synthetic data. However there are differences between the synthetic and original training data which affects the performance of the quantized network but none of the existing methods considers the differences. In this paper we propose an efficient data-free quantization via pseudo-label filtering which is the first to evaluate the synthetic data before quantization. We design a new metric for evaluating synthetic data using self-entropy which indicates the reliability of synthetic data. The synthetic data can be categorized with the metric into high- and low-reliable datasets for the following training process. Besides the multiple pseudo-labels are designed to label the synthetic data with different reliability which can provide valuable supervision information and avoid misleading training by low-reliable samples. Extensive experiments are implemented on several datasets including CIFAR-10 CIFAR-100 and ImageNet with various models. The experimental results show that our method can perform excellently and outperform existing methods in accuracy.
****
Revisiting Global Translation Estimation with Feature Tracks
Peilin Tao, Hainan Cui, Mengqi Rong, Shuhan Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20686-20696
Global translation estimation is a highly challenging step in the global structure from motion (SfM) algorithm. Many existing methods depend solely on relative translations leading to inaccuracies in low parallax scenes and degradation under collinear camera motion. While recent approaches aim to address these issues by incorporating feature tracks into objective functions they are often sensitive to outliers. In this paper we first revisit global translation estimation methods with feature tracks and categorize them into explicit and implicit methods. Then we highlight the superiority of the objective function based on the cross-product distance metric and propose a novel explicit global translation estimation framework that integrates both relative translations and feature tracks as input. To enhance the accuracy of input observations we re-estimate relative translations with the coplanarity constraint of the epipolar plane and propose a simple yet effective strategy to select reliable feature tracks. Finally the effectiveness of our approach is demonstrated through experiments on urban image sequences and unordered Internet images showcasing its superior accuracy and robustness compared to many state-of-the-art techniques.

****
Open-Set Domain Adaptation for Semantic Segmentation
Seun-An Choe, Ah-Hyung Shin, Keon-Hee Park, Jinwoo Choi, Gyeong-Moon Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23943-23953
Unsupervised domain adaptation (UDA) for semantic segmentation aims to transfer the pixel-wise knowledge from the labeled source domain to the unlabeled target domain. However current UDA methods typically assume a shared label space between source and target limiting their applicability in real-world scenarios where novel categories may emerge in the target domain. In this paper we introduce Open-Set Domain Adaptation for Semantic Segmentation (OSDA-SS) for the first time where the target domain includes unknown classes. We identify two major problems in the OSDA-SS scenario as follows: 1) the existing UDA methods struggle to predict the exact boundary of the unknown classes and 2) they fail to accurately predict the shape of the unknown classes. To address these issues we propose Boundary and Unknown Shape-Aware open-set domain adaptation coined BUS. Our BUS can accurately discern the boundaries between known and unknown classes in a contrastive manner using a novel dilation-erosion-based contrastive loss. In addition we propose OpenReMix a new domain mixing augmentation method that guides our model to effectively learn domain and size-invariant features for improving the shape detection of the known and unknown classes. Through extensive experiments we demonstrate that our proposed BUS effectively detects unknown classes in the challenging OSDA-SS scenario compared to the previous methods by a large margin.
****
Generative Powers of Ten
Xiaojuan Wang, Janne Kontkanen, Brian Curless, Steven M. Seitz, Ira Kemelmacher-Shlizerman, Ben Mildenhall, Pratul Srinivasan, Dor Verbin, Aleksander Holynski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7173-7182
We present a method that uses a text-to-image model to generate consistent content across multiple image scales enabling extreme semantic zooms into a scene e.g. ranging from a wide-angle landscape view of a forest to a macro shot of an insect sitting on one of the tree branches. We achieve this through a joint multi-scale diffusion sampling approach that encourages consistency across different scales while preserving the integrity of each individual sampling process. Since each generated scale is guided by a different text prompt our method enables deeper levels of zoom than traditional super-resolution methods that may struggle to create new contextual structure at vastly different scales. We compare our method qualitatively with alternative techniques in image super-resolution and outpainting and show that our method is most effective at generating consistent multi-scale content.
****
H-ViT: A Hierarchical Vision Transformer for Deformable Image Registration
Morteza Ghahremani, Mohammad Khateri, Bailiang Jian, Benedikt Wiestler, Ehsan Adeli, Christian Wachinger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11513-11523
This paper introduces a novel top-down representation approach for deformable image registration which estimates the deformation field by capturing various short- and long-range flow features at different scale levels. As a Hierarchical Vision Transformer (H-ViT) we propose a dual self-attention and cross-attention mechanism that uses high-level features in the deformation field to represent low-level ones enabling information streams in the deformation field across all voxel patch embeddings irrespective of their spatial proximity. Since high-level features contain abstract flow patterns such patterns are expected to effectively contribute to the representation of the deformation field in lower scales. When the self-attention module utilizes within-scale short-range patterns for representation the cross-attention modules dynamically look for the key tokens across different scales to further interact with the local query voxel patches. Our method shows superior accuracy and visual quality over the state-of-the-art registration methods in five publicly available datasets highlighting a substantial enhanc

ement in the performance of medical imaging registration. The project link is av ailable at https://mogvision.github.io/hvit.
****
Sculpting Holistic 3D Representation in Contrastive Language-Image-3D Pre-traini ng
Yipeng Gao, Zeyu Wang, Wei-Shi Zheng, Cihang Xie, Yuyin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp . 22998-23008
Contrastive learning has emerged as a promising paradigm for 3D open-world under standing i.e. aligning point cloud representation to image and text embedding sp ace individually. In this paper we introduce MixCon3D a simple yet effective met hod aiming to sculpt holistic 3D representation in contrastive language-image-3D pre-training. In contrast to point cloud only we develop the 3D object-level re presentation from complementary perspectives e.g. multi-view rendered images wit h the point cloud. Then MixCon3D performs language-3D contrastive learning compr ehensively depicting real-world 3D objects and bolstering text alignment. Additi onally we pioneer the first thorough investigation of various training recipes f or the 3D contrastive learning paradigm building a solid baseline with improved performance. Extensive experiments conducted on three representative benchmarks reveal that our method significantly improves over the baseline surpassing the p revious state-of-the-art performance on the challenging 1156-category Objaverse-LVIS dataset by 5.7%. The versatility of MixCon3D is showcased in applications s uch as text-to-3D retrieval and point cloud captioning further evidencing its ef ficacy in diverse scenarios. The code is available at https://github.com/UCSC-VL AA/MixCon3D.
****
Probing Synergistic High-Order Interaction in Infrared and Visible Image Fusion
Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, Feng Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2024, pp. 26384-26395
Infrared and visible image fusion aims to generate a fused image by integrating and distinguishing complementary information from multiple sources. While the cr oss-attention mechanism with global spatial interactions appears promising it on ly capture second-order spatial interactions neglecting higher-order interaction s in both spatial and channel dimensions. This limitation hampers the exploitati on of synergies between multi-modalities. To bridge this gap we introduce a Syne rgistic High-order Interaction Paradigm (SHIP) designed to systematically invest igate spatial fine-grained and global statistics collaborations between infrared and visible images across two fundamental dimensions: 1) Spatial dimension: we construct spatial fine-grained interactions through element-wise multiplication mathematically equivalent to global interactions and then foster high-order form ats by iteratively aggregating and evolving complementary information enhancing both efficiency and flexibility. 2) Channel dimension: expanding on channel inte ractions with first-order statistics (mean) we devise high-order channel interac tions to facilitate the discernment of inter-dependencies between source images based on global statistics. Harnessing high-order interactions significantly enh ances our model's ability to exploit multi-modal synergies leading in superior p erformance over state-of-the-art alternatives as shown through comprehensive exp eriments across various benchmarks.
****
VideoLLM-online: Online Video Large Language Model for Streaming Video
Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, J ia-Wei Liu, Ziteng Gao, Dongxing Mao, Mike Zheng Shou; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18407 -18418
Large Language Models (LLMs) have been enhanced with vision capabilities enablin g them to comprehend images videos and interleaved vision-language content. Howe ver the learning methods of these large multimodal models (LMMs) typically treat videos as predetermined clips rendering them less effective and efficient at ha ndling streaming video inputs. In this paper we propose a novel Learning-In-Vide

o-Stream (LIVE) framework which enables temporally aligned long-context and real-time dialogue within a continuous video stream. Our LIVE framework comprises comprehensive approaches to achieve video streaming dialogue encompassing: (1) a training objective designed to perform language modeling for continuous streaming inputs (2) a data generation scheme that converts offline temporal annotations into a streaming dialogue format and (3) an optimized inference pipeline to speed up interactive chat in real-world video streams. With our LIVE framework we develop a simplified model called VideoLLM-online and demonstrate its significant advantages in processing streaming videos. For instance our VideoLLM-online-7B model can operate at over 10 FPS on an A100 GPU for a 5-minute video clip from Ego4D narration. Moreover VideoLLM-online also showcases state-of-the-art performance on public offline video benchmarks such as recognition captioning and forecasting. The code model data and demo have been made available at showlab.github.io/videollm-online.
****
Text-conditional Attribute Alignment across Latent Spaces for 3D Controllable Face Image Synthesis
Feifan Xu, Rui Li, Si Wu, Yong Xu, Hau San Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9172-9181
With the advent of generative models and vision language pretraining significant improvement has been made in text-driven face manipulation. The text embedding can be used as target supervision for expression control.However it is non-trivial to associate with its 3D attributesi.e. pose and illumination. To address these issues we propose a Text-conditional Attribute aLignment approach for 3D controllable face image synthesis and our model is referred to as TcALign. Specifically since the 3D rendered image can be precisely controlled with the 3D face representation we first propose a Text-conditional 3D Editor to produce the target face representation to realize text-driven manipulation in the 3D space. An attribute embedding space spanned by the target-related attributes embeddings is also introduced to infer the disentangled task-specific direction. Next we train a cross-modal latent mapping network conditioned on the derived difference of 3D representation to infer a correct vector in the latent space of StyleGAN.Thiscorrection vector learning design can accurately transfer the attribute manipulation on 3D images to 2D images. We show that the proposed method delivers more precise text-driven multi-attribute manipulation for 3D controllable face image synthesis. Extensive qualitative and quantitative experiments verify the effectiveness and superiority of our method over the other competing methods.
****
ESCAPE: Encoding Super-keypoints for Category-Agnostic Pose Estimation
Khoi Duc Nguyen, Chen Li, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23491-23500
In this paper we tackle the task of category-agnostic pose estimation (CAPE) which aims to predict poses for objects of any category with few annotated samples. Previous works either rely on local matching between features of support and query samples or require support keypoint identifier. The former is prone to overfitting due to its sensitivity to sparse samples while the latter is impractical for the open-world nature of the task. To overcome these limitations we propose ESCAPE - a Bayesian framework that learns a prior over the features of keypoints. The prior can be expressed as a mixture of super-keypoints each being a high-level abstract keypoint that captures the statistics of semantically related keypoints from different categories. We estimate the super-keypoints from base categories and use them in adaptation to novel categories. The adaptation to an unseen category involves two steps: first we match each novel keypoint to a related super-keypoint; and second we transfer the knowledge encoded in the matched super-keypoints to the novel keypoints. For the first step we propose a learnable matching network to capture the relationship between the novel keypoints and the super-keypoints resulting in a more reliable matching. ESCAPE mitigates overfitting by directly transferring learned knowledge to novel categories while it does not use keypoint identifiers. We achieve state-of-the-art performance on the standard MP-100 benchmark.

****
Correcting Diffusion Generation through Resampling
Yujian Liu, Yang Zhang, Tommi Jaakkola, Shiyu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8713-8723
Despite diffusion models' superior capabilities in modeling complex distributions there are still non-trivial distributional discrepancies between generated and ground-truth images which has resulted in several notable problems in image generation including missing object errors in text-to-image generation and low image quality. Existing methods that attempt to address these problems mostly do not tend to address the fundamental cause behind these problems which is the distributional discrepancies and hence achieve sub-optimal results. In this paper we propose a particle filtering framework that can effectively address both problems by explicitly reducing the distributional discrepancies. Specifically our method relies on a set of external guidance including a small set of real images and a pre-trained object detector to gauge the distribution gap and then design the resampling weight accordingly to correct the gap. Experiments show that our methods can effectively correct missing object errors and improve image quality in various image generation tasks. Notably our method outperforms the existing strongest baseline by 5% in object occurrence and 1.0 in FID on MS-COCO. Our code is available at https://github.com/UCSB-NLP-Chang/diffusion_resampling.git.
****
Towards Better Vision-Inspired Vision-Language Models
Yun-Hao Cao, Kaixiang Ji, Ziyuan Huang, Chuanyang Zheng, Jiajia Liu, Jian Wang, Jingdong Chen, Ming Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13537-13547
Vision-language (VL) models have achieved unprecedented success recently in which the connection module is the key to bridge the modality gap. Nevertheless the abundant visual clues are not sufficiently exploited in most existing methods. On the vision side most existing approaches only use the last feature of the vision tower without using the low-level features. On the language side most existing methods only introduce shallow vision-language interactions. In this paper we present a vision-inspired vision-language connection module dubbed as VIVL which efficiently exploits the vision cue for VL models. To take advantage of the lowerlevel information from the vision tower a feature pyramid extractor (FPE) is introduced to combine features from different intermediate layers which enriches the visual cue with negligible parameters and computation overhead. To enhance VL interactions we propose deep vision-conditioned prompts (DVCP) that allows deep interactions of vision and language features efficiently. Our VIVL exceeds the previous state-of-the-art method by 18.1 CIDEr when training from scratch on the COCO caption task which greatly improves the data efficiency. When used as a plug-in module VIVL consistently improves the performance for various backbones and VL frameworks delivering new state-of-the-art results on multiple benchmarks e.g. NoCaps and VQAv2.
****
VSRD: Instance-Aware Volumetric Silhouette Rendering for Weakly Supervised 3D Object Detection
Zihua Liu, Hiroki Sakuma, Masatoshi Okutomi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17354-17363
Monocular 3D object detection poses a significant challenge in 3D scene understanding due to its inherently ill-posed nature in monocular depth estimation. Existing methods heavily rely on supervised learning using abundant 3D labels typically obtained through expensive and labor-intensive annotation on LiDAR point clouds. To tackle this problem we propose a novel weakly supervised 3D object detection framework named VSRD (Volumetric Silhouette Rendering for Detection) to train 3D object detectors without any 3D supervision but only weak 2D supervision. VSRD consists of multi-view 3D auto-labeling and subsequent training of monocular 3D object detectors using the pseudo labels generated in the auto-labeling stage. In the auto-labeling stage we represent the surface of each instance as a signed distance field (SDF) and render its silhouette as an instance mask through

our proposed instance-aware volumetric silhouette rendering. To directly optimiz
e the 3D bounding boxes through rendering we decompose the SDF of each instance
into the SDF of a cuboid and the residual distance field (RDF) that represents t
he residual from the cuboid. This mechanism enables us to optimize the 3D boundi
ng boxes in an end-to-end manner by comparing the rendered instance masks with t
he ground truth instance masks. The optimized 3D bounding boxes serve as effecti
ve training data for 3D object detection. We conduct extensive experiments on th
e KITTI-360 dataset demonstrating that our method outperforms the existing weakl
y supervised 3D object detection methods. The code is available at https://githu
b.com/skmhrk1209/VSRD.
****
RILA: Reflective and Imaginative Language Agent for Zero-Shot Semantic Audio-Vis
ual Navigation
Zeyuan Yang, Jiageng Liu, Peihao Chen, Anoop Cherian, Tim K. Marks, Jonathan Le
Roux, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2024, pp. 16251-16261
We leverage Large Language Models (LLM) for zeroshot Semantic Audio Visual Navig
ation (SAVN). Existing methods utilize extensive training demonstrations for rei
nforcement learning yet achieve relatively low success rates and lack generaliza
bility. The intermittent nature of auditory signals further poses additional obs
tacles to inferring the goal information. To address this challenge we present t
he Reflective and Imaginative Language Agent (RILA). By employing multi-modal mo
dels to process sensory data we instruct an LLM-based planner to actively explor
e the environment. During the exploration our agent adaptively evaluates and dis
misses inaccurate perceptual descriptions. Additionally we introduce an auxiliar
y LLMbased assistant to enhance global environmental comprehension by mapping ro
om layouts and providing strategic insights. Through comprehensive experiments a
nd analysis we show that our method outperforms relevant baselines without train
ing demonstrations from the environment and complementary semantic information.
****
Endow SAM with Keen Eyes: Temporal-spatial Prompt Learning for Video Camouflaged
 Object Detection
Wenjun Hui, Zhenfeng Zhu, Shuai Zheng, Yao Zhao; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19058-19067
The Segment Anything Model (SAM) a prompt-driven foundational model has demonstr
ated remarkable performance in natural image segmentation. However its applicati
on in video camouflaged object detection (VCOD) encounters challenges chiefly st
emming from the overlooked temporal-spatial associations and the unreliability o
f user-provided prompts for camouflaged objects that are difficult to discern wi
th the naked eye. To tackle the above issues we endow SAM with keen eyes and pro
pose the Temporal-spatial Prompt SAM (TSP-SAM) a novel approach tailored for VCO
D via an ingenious prompted learning scheme. Firstly motion-driven self-prompt l
earning is employed to capture the camouflaged object thereby bypassing the need
 for user-provided prompts. With the detected subtle motion cues across consecut
ive video frames the overall movement of the camouflaged object is captured for
more precise spatial localization. Subsequently to eliminate the prompt bias res
ulting from inter-frame discontinuities the long-range consistency within the vi
deo sequences is taken into account to promote the robustness of the self-prompt
s. It is also injected into the encoder of SAM to enhance the representational c
apabilities. Extensive experimental results on two benchmarks demonstrate that t
he proposed TSP-SAM achieves a significant improvement over the state-of-the-art
 methods. With the mIoU metric increasing by 7.8% and 9.6% TSP-SAM emerges as a
groundbreaking step forward in the field of VCOD.
****
TULIP: Multi-camera 3D Precision Assessment of Parkinson's Disease
Kyungdo Kim, Sihan Lyu, Sneha Mantri, Timothy W. Dunn; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22551
-22562
Parkinson's disease (PD) is a devastating movement disorder accelerating in glob
al prevalence but a lack of precision symptom measurement has made the developme

nt of effective therapies challenging. The Unified Parkinson's Disease Rating Scale (UPDRS) is the gold-standard for assessing motor symptom severity yet its manual scoring criteria are vague and subjective resulting in coarse and noisy clinical assessments. Machine learning approaches have the potential to modernize PD symptom assessments by making them more quantitative objective and scalable. However the lack of benchmark video datasets for PD motor exams hinders model development. Here we introduce the TULIP dataset to bridge this gap. TULIP emphasizes precision and comprehensiveness comprising multi-view video recordings (6 cameras) of all 25 UPDRS motor exam components together with ratings by 3 clinical experts in a cohort of Parkinson's patients and healthy controls. The multi-view recordings enable 3D reconstructions of body movement that better capture disease signatures than more conventional 2D methods. Using the dataset we establish a baseline model for predicting UPDRS scores from 3D poses illustrating how existing diagnostics could be automated. Looking ahead TULIP could aid the development of new precision diagnostics that transcend UPDRS scores providing a deeper understanding of PD and its potential treatments.
****
HybridNeRF: Efficient Neural Rendering via Adaptive Volumetric Surfaces
Haithem Turki, Vasu Agrawal, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Deva Ramanan, Michael Zollhöfer, Christian Richardt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19647-19656
Neural radiance fields provide state-of-the-art view synthesis quality but tend to be slow to render. One reason is that they make use of volume rendering thus requiring many samples (and model queries) per ray at render time. Although this representation is flexible and easy to optimize most real-world objects can be modeled more efficiently with surfaces instead of volumes requiring far fewer samples per ray. This observation has spurred considerable progress in surface representations such as signed distance functions but these may struggle to model semi-opaque and thin structures. We propose a method HybridNeRF that leverages the strengths of both representations by rendering most objects as surfaces while modeling the (typically) small fraction of challenging regions volumetrically. We evaluate HybridNeRF against the challenging Eyeful Tower dataset along with other commonly used view synthesis datasets. When comparing to state-of-the-art baselines including recent rasterization-based approaches we improve error rates by 15-30% while achieving real-time framerates (at least 36 FPS) for virtual-reality resolutions (2K x 2K).
****
AirPlanes: Accurate Plane Estimation via 3D-Consistent Embeddings
Jamie Watson, Filippo Aleotti, Mohamed Sayed, Zawar Qureshi, Oisin Mac Aodha, Gabriel Brostow, Michael Firman, Sara Vicente; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5270-5280
Extracting planes from a 3D scene is useful for downstream tasks in robotics and augmented reality. In this paper we tackle the problem of estimating the planar surfaces in a scene from posed images. Our first finding is that a surprisingly competitive baseline results from combining popular clustering algorithms with recent improvements in 3D geometry estimation. However such purely geometric methods are understandably oblivious to plane semantics which are crucial to discerning distinct planes. To overcome this limitation we propose a method that predicts multi-view consistent plane embeddings that complement geometry when clustering points into planes. We show through extensive evaluation on the ScanNetV2 dataset that our new method outperforms existing approaches and our strong geometric baseline for the task of plane estimation.
****
Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection
Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, Yao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10770-10780
In this paper we study the problem of generalizable synthetic image detection aiming to detect forgery images from diverse generative methods e.g. GANs and diff

usion models. Cutting-edge solutions start to explore the benefits of pre-traine
d models and mainly follow the fixed paradigm of solely training an attached cla
ssifier e.g. combining frozen CLIP-ViT with a learnable linear layer in UniFD. H
owever our analysis shows that such a fixed paradigm is prone to yield detectors
 with insufficient learning regarding forgery representations. We attribute the
key challenge to the lack of forgery adaptation and present a novel forgery-awar
e adaptive transformer approach namely FatFormer. Based on the pre-trained visio
n-language spaces of CLIP FatFormer introduces two core designs for the adaption
 to build generalized forgery representations. First motivated by the fact that
both image and frequency analysis are essential for synthetic image detection we
 develop a forgery-aware adapter to adapt image features to discern and integrat
e local forgery traces within image and frequency domains. Second we find that c
onsidering the contrastive objectives between adapted image features and text pr
ompt embeddings a previously overlooked aspect results in a nontrivial generaliz
ation improvement. Accordingly we introduce language-guided alignment to supervi
se the forgery adaptation with image and text prompts in FatFormer. Experiments
show that by coupling these two designs our approach tuned on 4-class ProGAN dat
a attains a remarkable detection performance achieving an average of 98% accurac
y to unseen GANs and surprisingly generalizes to unseen diffusion models with 95
% accuracy.
****
PostureHMR: Posture Transformation for 3D Human Mesh Recovery
Yu-Pei Song, Xiao Wu, Zhaoquan Yuan, Jian-Jun Qiao, Qiang Peng; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 9732-9741
Human Mesh Recovery (HMR) aims to estimate the 3D human body from 2D images whic
h is a challenging task due to inherent ambiguities in translating 2D observatio
ns to 3D space. A novel approach called PostureHMR is proposed to leverage a mul
ti-step diffusion-style process which converts this task into a posture transfor
mation from an SMPL T-pose mesh to the target mesh. To inject the learning proce
ss of posture transformation with the physical structure of the human body model
 a kinematics-based forward process is proposed to interpolate the intermediate
state with pose and shape decomposition. Moreover a mesh-to-posture (M2P) decode
r is designed by combining the input of 3D and 2D mesh constraints estimated fro
m the image to model the posture changes in the reverse process. It mitigates th
e difficulties of posture change learning directly from RGB pixels. To overcome
the limitation of pixel-level misalignment of modeling results with the input im
age a new trimap-based rendering loss is designed to highlight the areas with po
or recognition. Experiments conducted on three widely used datasets demonstrate
that the proposed approach outperforms the state-of-the-art methods.
****
Blur2Blur: Blur Conversion for Unsupervised Image Deblurring on Unknown Domains
Bang-Dang Pham, Phong Tran, Anh Tran, Cuong Pham, Rang Nguyen, Minh Hoai; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2024, pp. 2804-2813
This paper presents an innovative framework designed to train an image deblurrin
g algorithm tailored to a specific camera device. This algorithm works by transf
orming a blurry input image which is challenging to deblur into another blurry i
mage that is more amenable to deblurring. The transformation process from one bl
urry state to another leverages unpaired data consisting of sharp and blurry ima
ges captured by the target camera device. Learning this blur-to-blur transformat
ion is inherently simpler than direct blur-to-sharp conversion as it primarily i
nvolves modifying blur patterns rather than the intricate task of reconstructing
 fine image details. The efficacy of the proposed approach has been demonstrated
 through comprehensive experiments on various benchmarks where it significantly
outperforms state-of-the-art methods both quantitatively and qualitatively. Our
code and data are available at https://github.com/VinAIResearch/Blur2Blur
****
Dynamic Adapter Meets Prompt Tuning: Parameter-Efficient Transfer Learning for P
oint Cloud Analysis

Xin Zhou, Dingkang Liang, Wei Xu, Xingkui Zhu, Yihan Xu, Zhikang Zou, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14707-14717
Point cloud analysis has achieved outstanding performance by transferring point cloud pre-trained models. However existing methods for model adaptation usually update all model parameters i.e. full fine-tuning paradigm which is inefficient as it relies on high computational costs (e.g. training GPU memory) and massive storage space. In this paper we aim to study parameter-efficient transfer learning for point cloud analysis with an ideal trade-off between task performance and parameter efficiency. To achieve this goal we freeze the parameters of the default pre-trained models and then propose the Dynamic Adapter which generates a dynamic scale for each token considering the token significance to the downstream task. We further seamlessly integrate Dynamic Adapter with Prompt Tuning (DAPT) by constructing Internal Prompts capturing the instance-specific features for interaction. Extensive experiments conducted on five challenging datasets demonstrate that the proposed DAPT achieves superior performance compared to the full fine-tuning counterparts while significantly reducing the trainable parameters and training GPU memory by 95% and 35% respectively. Code is available at https://github.com/LMD0311/DAPT.
****
Exploring Vision Transformers for 3D Human Motion-Language Models with Motion Patches
Qing Yu, Mikihiro Tanaka, Kent Fujiwara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 937-946
To build a cross-modal latent space between 3D human motion and language acquiring large-scale and high-quality human motion data is crucial. However unlike the abundance of image data the scarcity of motion data has limited the performance of existing motion-language models. To counter this we introduce "motion patches" a new representation of motion sequences and propose using Vision Transformers (ViT) as motion encoders via transfer learning aiming to extract useful knowledge from the image domain and apply it to the motion domain. These motion patches created by dividing and sorting skeleton joints based on body parts in motion sequences are robust to varying skeleton structures and can be regarded as color image patches in ViT. We find that transfer learning with pre-trained weights of ViT obtained through training with 2D image data can boost the performance of motion analysis presenting a promising direction for addressing the issue of limited motion data. Our extensive experiments show that the proposed motion patches used jointly with ViT achieve state-of-the-art performance in the benchmarks of text-to-motion retrieval and other novel challenging tasks such as cross-skeleton recognition zero-shot motion classification and human interaction recognition which are currently impeded by the lack of data.
****
Motion-adaptive Separable Collaborative Filters for Blind Motion Deblurring
Chengxu Liu, Xuan Wang, Xiangyu Xu, Ruhao Tian, Shuai Li, Xueming Qian, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25595-25605
Eliminating image blur produced by various kinds of motion has been a challenging problem. Dominant approaches rely heavily on model capacity to remove blurring by reconstructing residual from blurry observation in feature space. These practices not only prevent the capture of spatially variable motion in the real world but also ignore the tailored handling of various motions in image space. In this paper we propose a novel real-world deblurring filtering model called the Motion-adaptive Separable Collaborative (MISC) Filter. In particular we use a motion estimation network to capture motion information from neighborhoods thereby adaptively estimating spatially-variant motion flow mask kernels weights and offsets to obtain the MISC Filter. The MISC Filter first aligns the motion-induced blurring patterns to the motion middle along the predicted flow direction and then collaboratively filters the aligned image through the predicted kernels weights and offsets to generate the output. This design can handle more generalized and complex motion in a spatially differentiated manner. Furthermore we analyze the

relationships between the motion estimation network and the residual reconstruction network. Extensive experiments on four widely used benchmarks demonstrate that our method provides an effective solution for real-world motion blur removal and achieves state-of-the-art performance. Code is available at https://github.com/ChengxuLiu/MISCFilter.
****
DART: Implicit Doppler Tomography for Radar Novel View Synthesis
Tianshu Huang, John Miller, Akarsh Prabhakara, Tao Jin, Tarana Laroia, Zico Kolter, Anthony Rowe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24118-24129
Simulation is an invaluable tool for radio-frequency system designers that enables rapid prototyping of various algorithms for imaging target detection classification and tracking. However simulating realistic radar scans is a challenging task that requires an accurate model of the scene radio frequency material properties and a corresponding radar synthesis function. Rather than specifying these models explicitly we propose DART - Doppler Aided Radar Tomography a Neural Radiance Field-inspired method which uses radar-specific physics to create a reflectance and transmittance-based rendering pipeline for range-Doppler images. We then evaluate DART by constructing a custom data collection platform and collecting a novel radar dataset together with accurate position and instantaneous velocity measurements from lidar-based localization. In comparison to state-of-the-art baselines DART synthesizes superior radar range-Doppler images from novel views across all datasets and additionally can be used to generate high quality tomographic images.
****
Wonder3D: Single Image to 3D using Cross-Domain Diffusion
Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9970-9980
In this work we introduce Wonder3D a novel method for generating high-fidelity textured meshes from single-view images with remarkable efficiency. Recent methods based on the Score Distillation Sampling (SDS) loss methods have shown the potential to recover 3D geometry from 2D diffusion priors but they typically suffer from time-consuming per-shape optimization and inconsistent geometry. In contrast certain works directly produce 3D information via fast network inferences but their results are often of low quality and lack geometric details. To holistically improve the quality consistency and efficiency of image-to-3D tasks we propose a cross-domain diffusion model that generates multi-view normal maps and the corresponding color images. To ensure consistency we employ a multi-view cross-domain attention mechanism that facilitates information exchange across views and modalities. Lastly we introduce a geometry-aware normal fusion algorithm that extracts high-quality surfaces from the multi-view 2D representations in only 2 3 minutes. Our extensive evaluations demonstrate that our method achieves high-quality reconstruction results robust generalization and remarkable efficiency compared to prior works.
****
Genuine Knowledge from Practice: Diffusion Test-Time Adaptation for Video Adverse Weather Removal
Yijun Yang, Hongtao Wu, Angelica I. Aviles-Rivero, Yulun Zhang, Jing Qin, Lei Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25606-25616
Real-world vision tasks frequently suffer from the appearance of unexpected adverse weather conditions including rain haze snow and raindrops. In the last decade convolutional neural networks and vision transformers have yielded outstanding results in single-weather video removal. However due to the absence of appropriate adaptation most of them fail to generalize to other weather conditions. Although ViWS-Net is proposed to remove adverse weather conditions in videos with a single set of pre-trained weights it is seriously blinded by seen weather at train-time and degenerates when coming to unseen weather during test-time. In this

work we introduce test-time adaptation into adverse weather removal in videos and propose the first framework that integrates test-time adaptation into the iterative diffusion reverse process. Specifically we devise a diffusion-based network with a novel temporal noise model to efficiently explore frame-correlated information in degraded video clips at training stage. During inference stage we introduce a proxy task named Diffusion Tubelet Self-Calibration to learn the primer distribution of test video stream and optimize the model by approximating the temporal noise model for online adaptation. Experimental results on benchmark datasets demonstrate that our Test-Time Adaptation method with Diffusion-based network(Diff-TTA) outperforms state-of-the-art methods in terms of restoring videos degraded by seen weather conditions. Its generalizable capability is validated with unseen weather conditions in synthesized and real-world videos.
****
Gradient-based Parameter Selection for Efficient Fine-Tuning
Zhi Zhang, Qizhe Zhang, Zijun Gao, Renrui Zhang, Ekaterina Shutova, Shiji Zhou, Shanghang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28566-28577
With the growing size of pre-trained models full fine-tuning and storing all the parameters for various downstream tasks is costly and infeasible. In this paper we propose a new parameter-efficient fine-tuning method Gradient-based Parameter Selection (GPS) demonstrating that only tuning a few selected parameters from the pre-trained model while keeping the remainder of the model frozen can generate similar or better performance compared with the full model fine-tuning method. Different from the existing popular and state-of-the-art parameter-efficient fine-tuning approaches our method does not introduce any additional parameters and computational costs during both the training and inference stages. Another advantage is the model-agnostic and non-destructive property which eliminates the need for any other design specific to a particular model. Compared with the full fine-tuning GPS achieves 3.33% (91.78% vs. 88.45% FGVC) and 9.61% (73.1% vs. 65.57% VTAB) improvement of the accuracy with tuning only 0.36% parameters of the pre-trained model on average over 24 image classification tasks; it also demonstrates a significant improvement of 17% and 16.8% in mDice and mIoU respectively on medical image segmentation task. Moreover GPS achieves state-of-the-art performance compared with existing PEFT methods. The code will be available in https://github.com/FightingFighting/GPS.git.
****
Clustering for Protein Representation Learning
Ruijie Quan, Wenguan Wang, Fan Ma, Hehe Fan, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 319-329
Protein representation learning is a challenging task that aims to capture the structure and function of proteins from their amino acid sequences. Previous methods largely ignored the fact that not all amino acids are equally important for protein folding and activity. In this article we propose a neural clustering framework that can automatically discover the critical components of a protein by considering both its primary and tertiary structure information. Our framework treats a protein as a graph where each node represents an amino acid and each edge represents a spatial or sequential connection between amino acids. We then apply an iterative clustering strategy to group the nodes into clusters based on their 1D and 3D positions and assign scores to each cluster. We select the highest-scoring clusters and use their medoid nodes for the next iteration of clustering until we obtain a hierarchical and informative representation of the protein. We evaluate on four protein-related tasks: protein fold classification enzyme reaction classification gene ontology term prediction and enzyme commission number prediction. Experimental results demonstrate that our method achieves state-of-the-art performance.
****
CorrMatch: Label Propagation via Correlation Matching for Semi-Supervised Semantic Segmentation
Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, Qibin Hou; Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 3097-3107
This paper presents a simple but performant semi-supervised semantic segmentatio
n approach called CorrMatch. Previous approaches mostly employ complicated train
ing strategies to leverage unlabeled data but overlook the role of correlation m
aps in modeling the relationships between pairs of locations. We observe that th
e correlation maps not only enable clustering pixels of the same category easily
 but also contain good shape information which previous works have omitted. Moti
vated by these we aim to improve the use efficiency of unlabeled data by designi
ng two novel label propagation strategies. First we propose to conduct pixel pro
pagation by modeling the pairwise similarities of pixels to spread the high-conf
idence pixels and dig out more. Then we perform region propagation to enhance th
e pseudo labels with accurate class-agnostic masks extracted from the correlatio
n maps. CorrMatch achieves great performance on popular segmentation benchmarks.
 Taking the DeepLabV3+ with ResNet-101 backbone as our segmentation model we rec
eive a 76%+ mIoU score on the Pascal VOC 2012 dataset with only 92 annotated ima
ges. Code is available at https://github.com/BBBBchan/CorrMatch .
****
Estimating Extreme 3D Image Rotations using Cascaded Attention
Shay Dekel, Yosi Keller, Martin Cadik; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2588-2598
Estimating large extreme inter-image rotations is critical for numerous computer
 vision domains involving images related by limited or non-overlapping fields of
 view. In this work we propose an attention-based approach with a pipeline of no
vel algorithmic components. First as rotation estimation pertains to image pairs
 we introduce an inter-image distillation scheme using Decoders to improve embed
dings. Second whereas contemporary methods compute a 4D correlation volume (4DCV
) encoding inter-image relationships we propose an Encoder-based cross-attention
 approach between activation maps to compute an enhanced equivalent of the 4DCV.
 Finally we present a cascaded Decoder-based technique for alternately refining
the cross-attention and the rotation query. Our approach outperforms current sta
te-of-the-art methods on extreme rotation estimation. We make our code publicly
available.
****
RichDreamer: A Generalizable Normal-Depth Diffusion Model for Detail Richness in
 Text-to-3D
Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao
 Yuan, Zilong Dong, Liefeng Bo, Xiaoguang Han; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9914-9925
Lifting 2D diffusion for 3D generation is a challenging problem due to the lack
of geometric prior and the complex entanglement of materials and lighting in nat
ural images. Existing methods have shown promise by first creating the geometry
through score-distillation sampling (SDS) applied to rendered surface normals fo
llowed by appearance modeling. However relying on a 2D RGB diffusion model to op
timize surface normals is suboptimal due to the distribution discrepancy between
 natural images and normals maps leading to instability in optimization. In this
 paper recognizing that the normal and depth information effectively describe sc
ene geometry and be automatically estimated from images we propose to learn a ge
neralizable Normal-Depth diffusion model for 3D generation. We achieve this by t
raining on the large-scale LAION dataset together with the generalizable image-t
o-depth and normal prior models. In an attempt to alleviate the mixed illuminati
on effects in the generated materials we introduce an albedo diffusion model to
impose data-driven constraints on the albedo component. Our experiments show tha
t when integrated into existing text-to-3D pipelines our models significantly en
hance the detail richness achieving state-of-the-art results. Our project page i
s at https://aigc3d.github.io/richdreamer/.
****
Adapt or Perish: Adaptive Sparse Transformer with Attentive Feature Refinement f
or Image Restoration
Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, Jufeng Yang; Proceedings o

f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2952-2963
Transformer-based approaches have achieved promising performance in image restoration tasks given their ability to model long-range dependencies which is crucial for recovering clear images. Though diverse efficient attention mechanism designs have addressed the intensive computations associated with using transformers they often involve redundant information and noisy interactions from irrelevant regions by considering all available tokens. In this work we propose an Adaptive Sparse Transformer (AST) to mitigate the noisy interactions of irrelevant areas and remove feature redundancy in both spatial and channel domains. AST comprises two core designs i.e. an Adaptive Sparse Self-Attention (ASSA) block and a Feature Refinement Feed-forward Network (FRFN). Specifically ASSA is adaptively computed using a two-branch paradigm where the sparse branch is introduced to filter out the negative impacts of low query-key matching scores for aggregating features while the dense one ensures sufficient information flow through the network for learning discriminative representations. Meanwhile FRFN employs an enhance-and-ease scheme to eliminate feature redundancy in channels enhancing the restoration of clear latent images. Experimental results on commonly used benchmarks have demonstrated the versatility and competitive performance of our method in several tasks including rain streak removal real haze removal and raindrop removal. The code and pre-trained models are available at https://github.com/joshyZhou/AST.
****
VINECS: Video-based Neural Character Skinning
Zhouyingcheng Liao, Vladislav Golyanik, Marc Habermann, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1377-1387
Rigging and skinning clothed human avatars is a challenging task and traditionally requires a lot of manual work and expertise. Recent methods addressing it either generalize across different characters or focus on capturing the dynamics of a single character observed under different pose configurations. However the former methods typically predict solely static skinning weights which perform poorly for highly articulated poses and the latter ones either require dense 3D character scans in different poses or cannot generate an explicit mesh with vertex correspondence over time. To address these challenges we propose a fully automated approach for creating a fully rigged character with pose-dependent skinning weights which can be solely learned from multi-view video. Therefore we first acquire a rigged template which is then statically skinned. Next a coordinate-based MLP learns a skinning weights field parameterized over the position in a canonical pose space and the respective pose. Moreover we introduce our pose- and view-dependent appearance field allowing us to differentiably render and supervise the posed mesh using multi-view imagery. We show that our approach outperforms state-of-the-art while not relying on dense 4D scans. More details can be found on our project page.
****
Zero-shot Referring Expression Comprehension via Structural Similarity Between Images and Captions
Zeyu Han, Fangrui Zhu, Qianru Lao, Huaizu Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14364-14374
Zero-shot referring expression comprehension aims at localizing bounding boxes in an image corresponding to provided textual prompts which requires: (i) a fine-grained disentanglement of complex visual scene and textual context and (ii) a capacity to understand relationships among disentangled entities. Unfortunately existing large vision-language alignment (VLA) models e.g. CLIP struggle with both aspects so cannot be directly used for this task. To mitigate this gap we leverage large foundation models to disentangle both images and texts into triplets in the format of (subject predicate object). After that grounding is accomplished by calculating the structural similarity matrix between visual and textual triplets with a VLA model and subsequently propagate it to an instance-level similarity matrix. Furthermore to equip VLA models with the ability of relationship un

derstanding we design a triplet-matching objective to fine-tune the VLA models on a collection of curated dataset containing abundant entity relationships. Experiments demonstrate that our visual grounding performance increase of up to 19.5% over the SOTA zero-shot model on RefCOCO/+/g. On the more challenging Who's Waldo dataset our zero-shot approach achieves comparable accuracy to the fully supervised model. Code is available at https://github.com/Show-han/Zeroshot_REC.
****
Domain Prompt Learning with Quaternion Networks
Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26637-26646
Prompt learning has emerged as an effective and data-efficient technique in large Vision-Language Models (VLMs). However when adapting VLMs to specialized domains such as remote sensing and medical imaging domain prompt learning remains underexplored. While large-scale domain-specific foundation models can help tackle this challenge their concentration on a single vision level makes it challenging to prompt both vision and language modalities. To overcome this we propose to leverage domain-specific knowledge from domain-specific foundation models to transfer the robust recognition ability of VLMs from generalized to specialized domains using quaternion networks. Specifically the proposed method involves using domain-specific vision features from domain-specific foundation models to guide the transformation of generalized contextual embeddings from the language branch into a specialized space within the quaternion networks. Moreover we present a hierarchical approach that generates vision prompt features by analyzing intermodal relationships between hierarchical language prompt features and domain-specific vision features. In this way quaternion networks can effectively mine the intermodal relationships in the specific domain facilitating domain-specific vision-language contrastive learning. Extensive experiments on domain-specific datasets show that our proposed method achieves new state-of-the-art results in prompt learning.
****
BEHAVIOR Vision Suite: Customizable Dataset Generation via Simulation
Yunhao Ge, Yihe Tang, Jiashu Xu, Cem Gokmen, Chengshu Li, Wensi Ai, Benjamin Jose Martinez, Arman Aydin, Mona Anvari, Ayush K Chakravarthy, Hong-Xing Yu, Josiah Wong, Sanjana Srivastava, Sharon Lee, Shengxin Zha, Laurent Itti, Yunzhu Li, Roberto Martín-Martín, Miao Liu, Pengchuan Zhang, Ruohan Zhang, Li Fei-Fei, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22401-22412
The systematic evaluation and understanding of computer vision models under varying conditions require large amounts of data with comprehensive and customized labels which real-world vision datasets rarely satisfy. While current synthetic data generators offer a promising alternative particularly for embodied AI tasks they often fall short for computer vision tasks due to low asset and rendering quality limited diversity and unrealistic physical properties. We introduce the BEHAVIOR Vision Suite (BVS) a set of tools and assets to generate fully customized synthetic data for systematic evaluation of computer vision models based on the newly developed embodied AI benchmark BEHAVIOR-1K. BVS supports a large number of adjustable parameters at the scene level (e.g. lighting object placement) the object level (e.g. joint configuration attributes such as "filled" and "folded") and the camera level (e.g. field of view focal length). Researchers can arbitrarily vary these parameters during data generation to perform controlled experiments. We showcase three example application scenarios: systematically evaluating the robustness of models across different continuous axes of domain shift evaluating scene understanding models on the same set of images and training and evaluating simulation-to-real transfer for a novel vision task: unary and binary state prediction. Project website: https://behavior-vision-suite.github.io/
****
Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers
Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, So

ng-Hai Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10324-10335

Recent advancements in 3D reconstruction from single images have been driven by the evolution of generative models. Prominent among these are methods based on Score Distillation Sampling (SDS) and the adaptation of diffusion models in the 3D domain. Despite their progress these techniques often face limitations due to slow optimization or rendering processes leading to extensive training and optimization times. In this paper we introduce a novel approach for single-view reconstruction that efficiently generates a 3D model from a single image via feed-forward inference. Our method utilizes two transformer-based networks namely a point decoder and a triplane decoder to reconstruct 3D objects using a hybrid Triplane-Gaussian intermediate representation. This hybrid representation strikes a balance achieving a faster rendering speed compared to implicit representations while simultaneously delivering superior rendering quality than explicit representations. The point decoder is designed for generating point clouds from single images offering an explicit representation which is then utilized by the triplane decoder to query Gaussian features for each point. This design choice addresses the challenges associated with directly regressing explicit 3D Gaussian attributes characterized by their non-structural nature. Subsequently the 3D Gaussians are decoded by an MLP to enable rapid rendering through splatting. Both decoders are built upon a scalable transformer-based architecture and have been efficiently trained on large-scale 3D datasets. The evaluations conducted on both synthetic datasets and real-world images demonstrate that our method not only achieves higher quality but also ensures a faster runtime in comparison to previous state-of-the-art techniques. Please see our project page at https://zouzx.github.io/TriplaneGaussian/

****

WateRF: Robust Watermarks in Radiance Fields for Protection of Copyrights

Youngdong Jang, Dong In Lee, MinHyuk Jang, Jong Wook Kim, Feng Yang, Sangpil Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12087-12097

The advances in the Neural Radiance Fields (NeRF) research offer extensive applications in diverse domains but protecting their copyrights has not yet been researched in depth. Recently NeRF watermarking has been considered one of the pivotal solutions for safely deploying NeRF-based 3D representations. However existing methods are designed to apply only to implicit or explicit NeRF representations. In this work we introduce an innovative watermarking method that can be employed in both representations of NeRF. This is achieved by fine-tuning NeRF to embed binary messages in the rendering process. In detail we propose utilizing the discrete wavelet transform in the NeRF space for watermarking. Furthermore we adopt a deferred back-propagation technique and introduce a combination with the patch-wise loss to improve rendering quality and bit accuracy with minimum trade-offs. We evaluate our method in three different aspects: capacity invisibility and robustness of the embedded watermarks in the 2D-rendered images. Our method achieves state-of-the-art performance with faster training speed over the compared state-of-the-art methods. Project page: https://kuai-lab.github.io/cvpr2024waterf/

****

Gaussian-Flow: 4D Reconstruction with Dynamic 3D Gaussian Particle

Youtian Lin, Zuozhuo Dai, Siyu Zhu, Yao Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21136-21145

We introduce Gaussian-Flow a novel point-based approach for fast dynamic scene reconstruction and real-time rendering from both multi-view and monocular videos. In contrast to the prevalent NeRF-based approaches hampered by slow training and rendering speeds our approach harnesses recent advancements in point-based 3D Gaussian Splatting (3DGS). Specifically a novel Dual-Domain Deformation Model (DDDM) is proposed to explicitly model attribute deformations of each Gaussian point where the time-dependent residual of each attribute is captured by a polynomial fitting in the time domain and a Fourier series fitting in the frequency domain. The proposed DDDM is capable of modeling complex scene deformations across l

ong video footage eliminating the need for training separate 3DGS for each frame or introducing an additional implicit neural field to model 3D dynamics. Moreover the explicit deformation modeling for discretized Gaussian points ensures ultra-fast training and rendering of a 4D scene which is comparable to the original 3DGS designed for static 3D reconstruction. Our proposed approach showcases a substantial efficiency improvement achieving a 5xfaster training speed compared to the per-frame 3DGS modeling. In addition quantitative results demonstrate that the proposed Gaussian-Flow significantly outperforms previous leading methods in novel view rendering quality.
****
Your Student is Better Than Expected: Adaptive Teacher-Student Collaboration for Text-Conditional Diffusion Models
Nikita Starodubcev, Dmitry Baranchuk, Artem Fedorov, Artem Babenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9275-9285
Knowledge distillation methods have recently shown to be a promising direction to speedup the synthesis of large-scale diffusion models by requiring only a few inference steps. While several powerful distillation methods were recently proposed the overall quality of student samples is typically lower compared to the teacher ones which hinders their practical usage. In this work we investigate the relative quality of samples produced by the teacher text-to-image diffusion model and its distilled student version. As our main empirical finding we discover that a noticeable portion of student samples exhibit superior fidelity compared to the teacher ones despite the approximate nature of the student. Based on this finding we propose an adaptive collaboration between student and teacher diffusion models for effective text-to-image synthesis. Specifically the distilled model produces an initial image sample and then an oracle decides whether it needs further improvements with the teacher model. Extensive experiments demonstrate that the designed pipeline surpasses state-of-the-art text-to-image alternatives for various inference budgets in terms of human preference. Furthermore the proposed approach can be naturally used in popular applications such as text-guided image editing and controllable generation.
****
DiVAS: Video and Audio Synchronization with Dynamic Frame Rates
Clara Fernandez-Labrador, Mertcan Akçay, Eitan Abecassis, Joan Massich, Christopher Schroers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26846-26854
Synchronization issues between audio and video are one of the most disturbing quality defects in film production and live broadcasting. Even a discrepancy as short as 45 millisecond can degrade the viewer's experience enough to warrant manual quality checks over entire movies. In this paper we study the automatic discovery of such issues. Specifically we focus on the alignment of lip movements with spoken words targeting realistic production scenarios which can include background noise and music intricate head poses excessive makeup or scenes with multiple individuals where the speaker is unknown. Our model's robustness also extends to various media specifications including different video frame rates and audio sample rates. To address these challenges we present a model fully based on transformers that encodes face crops or full video frames and raw audio using timestamp information identifies the speaker and provides highly accurate synchronization predictions much faster than previous methods.
****
SHViT: Single-Head Vision Transformer with Memory Efficient Macro Design
Seokju Yun, Youngmin Ro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5756-5767
Recently efficient Vision Transformers have shown great performance with low latency on resource-constrained devices. Conventionally they use 4x4 patch embeddings and a 4-stage structure at the macro level while utilizing sophisticated attention with multi-head configuration at the micro level. This paper aims to address computational redundancy at all design levels in a memory-efficient manner. We discover that using larger-stride patchify stem not only reduces memory access

costs but also achieves competitive performance by leveraging token representations with reduced spatial redundancy from the early stages. Furthermore our preliminary analyses suggest that attention layers in the early stages can be substituted with convolutions and several attention heads in the latter stages are computationally redundant. To handle this we introduce a single-head attention module that inherently prevents head redundancy and simultaneously boosts accuracy by parallelly combining global and local information. Building upon our solutions we introduce SHViT a Single-Head Vision Transformer that obtains the state-of-the-art speed-accuracy tradeoff. For example on ImageNet-1k our SHViT-S4 is 3.3x 8.1x and 2.4x faster than MobileViTv2x1.0 on GPU CPU and iPhone12 mobile device respectively while being 1.3% more accurate. For object detection and instance segmentation on MS COCO using Mask-RCNN head our model achieves performance comparable to FastViT-SA12 while exhibiting 3.8x and 2.0x lower backbone latency on GPU and mobile device respectively.
****
HDRFlow: Real-Time HDR Video Reconstruction with Large Motions
Gangwei Xu, Yujin Wang, Jinwei Gu, Tianfan Xue, Xin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24851-24860
Reconstructing High Dynamic Range (HDR) video from image sequences captured with alternating exposures is challenging especially in the presence of large camera or object motion. Existing methods typically align low dynamic range sequences using optical flow or attention mechanism for deghosting. However they often struggle to handle large complex motions and are computationally expensive. To address these challenges we propose a robust and efficient flow estimator tailored for real-time HDR video reconstruction named HDRFlow. HDRFlow has three novel designs: an HDR-domain alignment loss (HALoss) an efficient flow network with a multi-size large kernel (MLK) and a new HDR flow training scheme. The HALoss supervises our flow network to learn an HDR-oriented flow for accurate alignment in saturated and dark regions. The MLK can effectively model large motions at a negligible cost. In addition we incorporate synthetic data Sintel into our training dataset utilizing both its provided forward flow and backward flow generated by us to supervise our flow network enhancing our performance in large motion regions. Extensive experiments demonstrate that our HDRFlow outperforms previous methods on standard benchmarks. To the best of our knowledge HDRFlow is the first real-time HDR video reconstruction method for video sequences captured with alternating exposures capable of processing 720p resolution inputs at 25ms.
****
SPIDeRS: Structured Polarization for Invisible Depth and Reflectance Sensing
Tomoki Ichikawa, Shohei Nobuhara, Ko Nishino; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25077-25085
Can we capture shape and reflectance in stealth? Such capability would be valuable for many application domains in vision xR robotics and HCI. We introduce structured polarization for invisible depth and reflectance sensing (SPIDeRS) the first depth and reflectance sensing method using patterns of polarized light. The key idea is to modulate the angle of linear polarization (AoLP) of projected light at each pixel. The use of polarization makes it invisible and lets us recover not only depth but also directly surface normals and even reflectance. We implement SPIDeRS with a liquid crystal spatial light modulator (SLM) and a polarimetric camera. We derive a novel method for robustly extracting the projected structured polarization pattern from the polarimetric object appearance. We evaluate the effectiveness of SPIDeRS by applying it to a number of real-world objects. The results show that our method successfully reconstructs object shapes of various materials and is robust to diffuse reflection and ambient light. We also demonstrate relighting using recovered surface normals and reflectance. We believe SPIDeRS opens a new avenue of polarization use in visual sensing.
****
SuperNormal: Neural Surface Reconstruction via Multi-View Normal Integration
Xu Cao, Takafumi Taketomi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20581-20590

We present SuperNormal a fast high-fidelity approach to multi-view 3D reconstruc
tion using surface normal maps. With a few minutes SuperNormal produces detailed
 surfaces on par with 3D scanners. We harness volume rendering to optimize a neu
ral signed distance function (SDF) powered by multi-resolution hash encoding. To
 accelerate training we propose directional finite difference and patchbased ray
 marching to approximate the SDF gradients numerically. While not compromising r
econstruction quality this strategy is nearly twice as efficient as analytical g
radients and about three times faster than axis-aligned finite difference. Exper
iments on the benchmark dataset demonstrate the superiority of SuperNormal in ef
ficiency and accuracy compared to existing multi-view photometric stereo methods
. On our captured objects SuperNormal produces more fine-grained geometry than r
ecent neural 3D reconstruction methods. Our code is available at https://github.
com/CyberAgentAILab/SuperNormal.git.
****
Instance-aware Contrastive Learning for Occluded Human Mesh Reconstruction
Mi-Gyeong Gwon, Gi-Mun Um, Won-Sik Cheong, Wonjun Kim; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10553
-10562
A simple yet effective method for occlusion-robust 3D human mesh reconstruction
from a single image is presented in this paper. Although many recent studies hav
e shown the remarkable improvement in human mesh reconstruction it is still diff
icult to generate accurate meshes when person-to-person occlusion occurs due to
the ambiguity of who a body part belongs to. To address this problem we propose
an instance-aware contrastive learning scheme. Specifically joint features belon
ging to the target human are trained to be proximate with the anchor feature (i.
e. feature extracted from the body center position). On the other hand anchor fe
atures of different human instances are forced to be far apart so that joint fea
tures of each person can be clearly distinguished from others. By interpreting t
he joint possession based on such contrastive learning scheme the proposed metho
d easily understands the spatial occupancy of body parts for each person in a gi
ven image thus can reconstruct reliable human meshes even with severely overlapp
ed cases between multiple persons. Experimental results on benchmark datasets de
monstrate the robustness of the proposed method compared to previous approaches
under person-to-person occlusions. The code and model are publicly available at:
 https://github.com/DCVL-3D/InstanceHMR_release.
****
ADFactory: An Effective Framework for Generalizing Optical Flow with NeRF
Han Ling, Quansen Sun, Yinghui Sun, Xian Xu, Xinfeng Li; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 205
91-20600
A significant challenge facing current optical flow methods is the difficulty in
 generalizing them well to the real world. This is mainly due to the lack of lar
ge-scale real-world datasets and existing self-supervised methods are limited by
 indirect loss and occlusions resulting in fuzzy outcomes. To address this chall
enge we introduce a novel optical flow training framework: automatic data factor
y (ADF). ADF only requires RGB images as input to effectively train the optical
flow network on the target data domain. Specifically we use advanced NeRF techno
logy to reconstruct scenes from photo groups collected by a monocular camera and
 then calculate optical flow labels between camera pose pairs based on the rende
ring results. To eliminate erroneous labels caused by defects in the scene recon
structed by NeRF we screened the generated labels from multiple aspects such as
optical flow matching accuracy radiation field confidence and depth consistency.
 The filtered labels can be directly used for network supervision. Experimentall
y the generalization ability of ADF on KITTI surpasses existing self-supervised
optical flow and monocular scene flow algorithms. In addition ADF achieves impre
ssive results in real-world zero-point generalization evaluations and surpasses
most supervised methods.
****
Robust Noisy Correspondence Learning with Equivariant Similarity Consistency
Yuchen Yang, Likai Wang, Erkun Yang, Cheng Deng; Proceedings of the IEEE/CVF Con

ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17700-17709
The surge in multi-modal data has propelled cross-modal matching to the forefron
t of research interest. However the challenge lies in the laborious and expensiv
e process of curating a large and accurately matched multimodal dataset. Commonl
y sourced from the Internet these datasets often suffer from a significant prese
nce of mismatched data impairing the performance of matching models. To address
this problem we introduce a novel regularization approach named Equivariant Simi
larity Consistency (ESC) which can facilitate robust clean and noisy data separa
tion and improve the training for cross-modal matching. Intuitively our method p
osits that the semantic variations caused by image changes should be proportiona
l to those caused by text changes for any two matched samples. Accordingly we fi
rst calculate the ESC by comparing image and text semantic variations between a
set of elaborated anchor points and other undivided training data. Then pairs wi
th high ESC are filtered out as noisy correspondence pairs. We implement our met
hod by combining the ESC with a traditional hinge-based triplet loss. Extensive
experiments on three widely used datasets including Flickr30K MS-COCO and Concep
tual Captions verify the effectiveness of our method.
****
CommonCanvas: Open Diffusion Models Trained on Creative-Commons Images
Aaron Gokaslan, A. Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson
, Mihir Patel, Jonathan Frankle, Cory Stephenson, Volodymyr Kuleshov; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 8250-8260
We train a set of open text-to-image (T2I) diffusion models on a dataset of cura
ted Creative-Commons-licensed (CC) images which yields models that are competiti
ve with Stable Diffusion 2 (SD2). This task presents two challenges: (1) high-re
solution CC images lack the captions necessary to train T2I models; (2) CC image
s are relatively scarce. To address these challenges we use an intuitive transfe
r learning technique to produce a set of high-quality synthetic captions paired
with our assembled CC images. We then develop a data- and compute-efficient trai
ning recipe that requires as little as 3% of the LAION data (i.e. roughly 70 mil
lion examples) needed to train existing SD2 models but obtains the same quality.
These results indicate that we have a sufficient number of CC images (also roug
hly 70 million) for training high-quality models. Our recipe also implements a v
ariety of optimizations that achieve 2.71x training speed-ups enabling rapid mod
el iteration. We leverage this recipe to train several high-quality T2I mod- els
which we dub the CommonCanvas family. Our largest model achieves comparable per
formance to SD2 on human evaluation even though we use a synthetically captioned
CC-image dataset that is only <3% the size of LAION for training. We release ou
r models data and code on GitHub.
****
Prompt-Driven Referring Image Segmentation with Instance Contrasting
Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng, Hongliang Li; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 4124-4134
Referring image segmentation (RIS) aims to segment the target referent described
by natural language. Recently large-scale pre-trained models e.g. CLIP and SAM
have been successfully applied in many downstream tasks but they are not well ad
apted to RIS task due to inter-task differences. In this paper we propose a new
prompt-driven framework named Prompt-RIS which bridges CLIP and SAM end-to-end a
nd transfers their rich knowledge and powerful capabilities to RIS task through
prompt learning. To adapt CLIP to pixel-level task we first propose a Cross-Moda
l Prompting method which acquires more comprehensive vision-language interaction
and fine-grained text-to-pixel alignment by performing bidirectional prompting.
Then the prompt-tuned CLIP generates masks points and text prompts for SAM to g
enerate more accurate mask predictions. Moreover we further propose Instance Con
trastive Learning to improve the model's discriminability to different instances
and robustness to diverse languages describing the same instance. Extensive exp
eriments demonstrate that the performance of our method outperforms the state-of
-the-art methods consistently in both general and open-vocabulary settings.

****
Image Sculpting: Precise Object Editing with 3D Geometry Control

Jiraphon Yenphraphai, Xichen Pan, Sainan Liu, Daniele Panozzo, Saining Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4241-4251

We present Image Sculpting a new framework for editing 2D images by incorporating tools from 3D geometry and graphics. This approach differs markedly from existing methods which are confined to 2D spaces and typically rely on textual instructions leading to ambiguity and limited control. Image Sculpting converts 2D objects into 3D enabling direct interaction with their 3D geometry. Post-editing these objects are re-rendered into 2D merging into the original image to produce high-fidelity results through a coarse-to-fine enhancement process. The framework supports precise quantifiable and physically-plausible editing options such as pose editing rotation translation 3D composition carving and serial addition. It marks an initial step towards combining the creative freedom of generative models with the precision of graphics pipelines.

****
Compositional Video Understanding with Spatiotemporal Structure-based Transformers

Hoyeoung Yun, Jinwoo Ahn, Minseo Kim, Eun-Sol Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18751-18760

In this paper we suggest a new novel method to understand complex semantic structures through long video inputs. Conventional methods for understanding videos have been focused on short-term clips and trained to get visual representations for the short clips using convolutional neural networks or transformer architectures. However most real-world videos are composed of long videos ranging from minutes to hours therefore it essentially brings limitations to understanding the overall semantic structures of the long videos by dividing them into small clips and learning the representations of them. We suggest a new algorithm to learn the multi-granular semantic structures of videos by defining spatiotemporal high-order relationships among object-based representations as semantic units. The proposed method includes a new transformer architecture capable of learning spatiotemporal graphs and a compositional learning method to learn disentangled features for each semantic unit. Using the suggested method we resolve the challenging video task which is compositional generalization understanding of unseen videos. In experiments we demonstrate new state-of-the-art performances for two challenging video datasets.

****
3D LiDAR Mapping in Dynamic Environments using a 4D Implicit Neural Representation

Xingguang Zhong, Yue Pan, Cyrill Stachniss, Jens Behley; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15417-15427

Building accurate maps is a key building block to enable reliable localization planning and navigation of autonomous vehicles. We propose a novel approach for building accurate 3D maps of dynamic environments utilizing a sequence of LiDAR scans. To this end we propose encoding the 4D scene into a novel spatio-temporal implicit neural map representation by fitting a time-dependent truncated signed distance function to each point. Using our representation we can extract the static map by filtering the dynamic parts. Our neural representation is based on sparse feature grids a globally shared decoder and time-dependent basis functions which can be jointly optimized in an unsupervised fashion. To learn this representation from a sequence of LiDAR scans we design a simple yet efficient loss function to supervise the map optimization in a piecewise way. We evaluate our approach on various scenes containing moving objects in terms of the reconstruction quality of static maps and the segmentation of dynamic point clouds. The experimental results demonstrate that our method is capable of removing the dynamic part of the input point clouds while reconstructing accurate and complete large-scale 3D maps outperforming several state-of-the-art methods for static map generat

ion and scene reconstruction.
****
What When and Where? Self-Supervised Spatio-Temporal Grounding in Untrimmed Multi-Action Videos from Narrated Instructions
Brian Chen, Nina Shvetsova, Andrew Rouditchenko, Daniel Kondermann, Samuel Thomas, Shih-Fu Chang, Rogerio Feris, James Glass, Hilde Kuehne; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18419-18429
Spatio-temporal grounding describes the task of localizing events in space and time e.g. in video data based on verbal descriptions only. Models for this task are usually trained with human-annotated sentences and bounding box supervision. This work addresses this task from a multimodal supervision perspective proposing a framework for spatio-temporal action grounding trained on loose video and subtitle supervision only without human annotation. To this end we combine local representation learning which focuses on leveraging fine-grained spatial information with a global representation encoding that captures higher-level representations and incorporates both in a joint approach. To evaluate this challenging task in a real-life setting a new benchmark dataset is proposed providing dense spatio-temporal grounding annotations in long untrimmed multi-action instructional videos for over 5K events. We evaluate the proposed approach and other methods on the proposed and standard downstream tasks showing that our method improves over current baselines in various settings including spatial temporal and untrimmed multi-action spatio-temporal grounding.
****
FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects
Bowen Wen, Wei Yang, Jan Kautz, Stan Birchfield; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17868-17879
We present FoundationPose a unified foundation model for 6D object pose estimation and tracking supporting both model-based and model-free setups. Our approach can be instantly applied at test-time to a novel object without finetuning as long as its CAD model is given or a small number of reference images are captured. Thanks to the unified framework the downstream pose estimation modules are the same in both setups with a neural implicit representation used for efficient novel view synthesis when no CAD model is available. Strong generalizability is achieved via large-scale synthetic training aided by a large language model (LLM) a novel transformer-based architecture and contrastive learning formulation. Extensive evaluation on multiple public datasets involving challenging scenarios and objects indicate our unified approach outperforms existing methods specialized for each task by a large margin. In addition it even achieves comparable results to instance-level methods despite the reduced assumptions. Project page: https://nvlabs.github.io/FoundationPose/
****
How Far Can We Compress Instant-NGP-Based NeRF?
Yihang Chen, Qianyi Wu, Mehrtash Harandi, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20321-20330
In recent years Neural Radiance Field (NeRF) has demonstrated remarkable capabilities in representing 3D scenes. To expedite the rendering process learnable explicit representations have been introduced for combination with implicit NeRF representation which however results in a large storage space requirement. In this paper we introduce the Context-based NeRF Compression (CNC) framework which leverages highly efficient context models to provide a storage-friendly NeRF representation. Specifically we excavate both level-wise and dimension-wise context dependencies to enable probability prediction for information entropy reduction. Additionally we exploit hash collision and occupancy grids as strong prior knowledge for better context modeling. To the best of our knowledge we are the first to construct and exploit context models for NeRF compression. We achieve a size reduction of 100X and 70X with improved fidelity against the baseline Instant-NGP on Synthesic-NeRF and Tanks and Temples datasets respectively. Additionally we attain 86.7% and 82.3% storage size reduction against the SOTA NeRF compression

method BiRF. Our code is available here: https://github.com/YihangChen-ee/CNC.
****
PFStorer: Personalized Face Restoration and Super-Resolution
Tuomas Varanka, Tapani Toivonen, Soumya Tripathy, Guoying Zhao, Erman Acar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2372-2381
Recent developments in face restoration have achieved remarkable results in producing high-quality and lifelike outputs. The stunning results however often fail to be faithful with respect to the identity of the person as the models lack necessary context. In this paper we explore the potential of personalized face restoration with diffusion models. In our approach a restoration model is personalized using a few images of the identity leading to tailored restoration with respect to the identity while retaining fine-grained details. By using independent trainable blocks for personalization the rich prior of a base restoration model can be exploited to its fullest. To avoid the model relying on parts of identity left in the conditioning low-quality images a generative regularizer is employed. With a learnable parameter the model learns to balance between the details generated based on the input image and the degree of personalization. Moreover we improve the training pipeline of face restoration models to enable an alignment-free approach. We showcase the robust capabilities of our approach in several real-world scenarios with multiple identities demonstrating our method's ability to generate fine-grained details with faithful restoration. In the user study we evaluate the perceptual quality and faithfulness of the generated details with our method being voted best 61% of the time compared to the second best with 25% of the votes.
****
TextureDreamer: Image-Guided Texture Synthesis Through Geometry-Aware Diffusion
Yu-Ying Yeh, Jia-Bin Huang, Changil Kim, Lei Xiao, Thu Nguyen-Phuoc, Numair Khan, Cheng Zhang, Manmohan Chandraker, Carl S Marshall, Zhao Dong, Zhengqin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4304-4314
We present TextureDreamer a novel image-guided texture synthesis method to transfer relightable textures from a small number of input images (3 to 5) to target 3D shapes across arbitrary categories. Texture creation is a pivotal challenge in vision and graphics. Industrial companies hire experienced artists to manually craft textures for 3D assets. Classical methods require densely sampled views and accurately aligned geometry while learning-based methods are confined to category-specific shapes within the dataset. In contrast TextureDreamer can transfer highly detailed intricate textures from real-world environments to arbitrary objects with only a few casually captured images potentially significantly democratizing texture creation. Our core idea personalized geometry-aware score distillation (PGSD) draws inspiration from recent advancements in diffuse models including personalized modeling for texture information extraction score distillation for detailed appearance synthesis and explicit geometry guidance with ControlNet. Our integration and several essential modifications substantially improve the texture quality. Experiments on real images spanning different categories show that TextureDreamer can successfully transfer highly realistic semantic meaningful texture to arbitrary objects surpassing the visual quality of previous state-of-the-art. Project page: https://texturedreamer.github.io
****
Boosting Image Quality Assessment through Efficient Transformer Adaptation with Local Feature Enhancement
Kangmin Xu, Liang Liao, Jing Xiao, Chaofeng Chen, Haoning Wu, Qiong Yan, Weisi Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2662-2672
Image Quality Assessment (IQA) constitutes a fundamental task within the field of computer vision yet it remains an unresolved challenge owing to the intricate distortion conditions diverse image contents and limited availability of data. Recently the community has witnessed the emergence of numerous large-scale pretrained foundation models. However it remains an open problem whether the scaling l

aw in high-level tasks is also applicable to IQA tasks which are closely related to low-level clues. In this paper we demonstrate that with a proper injection of local distortion features a larger pretrained vision transformer (ViT) foundation model performs better in IQA tasks. Specifically for the lack of local distortion structure and inductive bias of the large-scale pretrained ViT we use another pretrained convolution neural networks (CNNs) which is well known for capturing the local structure to extract multi-scale image features. Further we propose a local distortion extractor to obtain local distortion features from the pretrained CNNs and a local distortion injector to inject the local distortion features into ViT. By only training the extractor and injector our method can benefit from the rich knowledge in the powerful foundation models and achieve state-of-the-art performance on popular IQA datasets indicating that IQA is not only a low-level problem but also benefits from stronger high-level features drawn from large-scale pretrained models. Codes are publicly available at: https://github.com/NeosXu/LoDa.
****
Hyperbolic Anomaly Detection
Huimin Li, Zhentao Chen, Yunhao Xu, Junlin Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17511-17520
Anomaly detection is a challenging computer vision task in industrial scenario. Advancements in deep learning constantly revolutionize vision-based anomaly detection methods and considerable progress has been made in both supervised and self-supervised anomaly detection. The commonly-used pipeline is to optimize the model by constraining the feature embeddings using a distance-based loss function. However these methods work in Euclidean space and they cannot well exploit the data lied in non-Euclidean space. In this paper we are the first to explore anomaly detection task in hyperbolic space that is a representative of non-Euclidean space and propose a hyperbolic anomaly detection (HypAD) method. Specifically we first extract image features and then map them from Euclidean space to hyperbolic space where the hyperbolic distance metric is employed to optimize the proposed HypAD. Extensive experiments on the benchmarking datasets including MVTec AD and VisA show that our HypAD approach obtains the state-of-the-art performance demonstrating the effectiveness of our HypAD and the promise of investigating anomaly detection in hyperbolic space.
****
VLP: Vision Language Planning for Autonomous Driving
Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, Liu Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14760-14769
Autonomous driving is a complex and challenging task that aims at safe motion planning through scene understanding and reasoning. While vision-only autonomous driving methods have recently achieved notable performance through enhanced scene understanding several key issues including lack of reasoning low generalization performance and long-tail scenarios still need to be addressed. In this paper we present VLP a novel Vision-Language-Planning framework that exploits language models to bridge the gap between linguistic understanding and autonomous driving. VLP enhances autonomous driving systems by strengthening both the source memory foundation and the self-driving car's contextual understanding. VLP achieves state-of-the-art end-to-end planning performance on the challenging NuScenes dataset by achieving 35.9% and 60.5% reduction in terms of average L2 error and collision rates respectively compared to the previous best method. Moreover VLP shows improved performance in challenging long-tail scenarios and strong generalization capabilities when faced with new urban environments.
****
Attention Calibration for Disentangled Text-to-Image Personalization
Yanbing Zhang, Mengping Yang, Qin Zhou, Zhe Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4764-4774
Recent thrilling progress in large-scale text-to-image (T2I) models has unlocked unprecedented synthesis quality of AI-generated content (AIGC) including image generation 3D and video composition. Further personalized techniques enable appe

aling customized production of a novel concept given only several images as refe
rence. However an intriguing problem persists: Is it possible to capture multipl
e novel concepts from one single reference image? In this paper we identify that
 existing approaches fail to preserve visual consistency with the reference imag
e and eliminate cross-influence from concepts. To alleviate this we propose an a
ttention calibration mechanism to improve the concept-level understanding of the
 T2I model. Specifically we first introduce new learnable modifiers bound with c
lasses to capture attributes of multiple concepts. Then the classes are separate
d and strengthened following the activation of the cross-attention operation ens
uring comprehensive and self-contained concepts. Additionally we suppress the at
tention activation of different classes to mitigate mutual influence among conce
pts. Together our proposed method dubbed DisenDiff can learn disentangled multip
le concepts from one single image and produce novel customized images with learn
ed concepts. We demonstrate that our method outperforms the current state of the
 art in both qualitative and quantitative evaluations. More importantly our prop
osed techniques are compatible with LoRA and inpainting pipelines enabling more
interactive experiences.
****
ProMark: Proactive Diffusion Watermarking for Causal Attribution
Vishal Asnani, John Collomosse, Tu Bui, Xiaoming Liu, Shruti Agarwal; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 10802-10811
Generative AI (GenAI) is transforming creative workflows through the capability
to synthesize and manipulate images via high-level prompts. Yet creatives are no
t well supported to receive recognition or reward for the use of their content i
n GenAI training. To this end we propose ProMark a causal attribution technique
to attribute a synthetically generated image to its training data concepts like
objects motifs templates artists or styles. The concept information is proactive
ly embedded into the input training images using imperceptible watermarks and th
e diffusion models (unconditional or conditional) are trained to retain the corr
esponding watermarks in generated images. We show that we can embed as many as 2
^ 16 unique watermarks into the training data and each training image can conta
in more than one watermark. ProMark can maintain image quality whilst outperform
ing correlation-based attribution. Finally several qualitative examples are pres
ented providing the confidence that the presence of the watermark conveys a caus
ative relationship between training data and synthetic images.
****
One-Shot Structure-Aware Stylized Image Synthesis
Hansam Cho, Jonghyun Lee, Seunggyu Chang, Yonghyun Jeong; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 83
02-8311
While GAN-based models have been successful in image stylization tasks they ofte
n struggle with structure preservation while stylizing a wide range of input ima
ges. Recently diffusion models have been adopted for image stylization but still
 lack the capability to maintain the original quality of input images. Building
on this we propose OSASIS: a novel one-shot stylization method that is robust in
 structure preservation. We show that OSASIS is able to effectively disentangle
the semantics from the structure of an image allowing it to control the level of
 content and style implemented to a given input. We apply OSASIS to various expe
rimental settings including stylization with out-of-domain reference images and
stylization with text-driven manipulation. Results show that OSASIS outperforms
other stylization methods especially for input images that were rarely encounter
ed during training providing a promising solution to stylization via diffusion m
odels.
****
GPT4Point: A Unified Framework for Point-Language Understanding and Generation
Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, He
ngshuang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2024, pp. 26417-26427
Multimodal Large Language Models (MLLMs) have excelled in 2D image-text comprehe

nsion and image generation but their understanding of the 3D world is notably de
ficient limiting progress in 3D language understanding and generation. To solve
this problem we introduce GPT4Point an innovative groundbreaking point-language
multimodal model designed specifically for unified 3D object understanding and g
eneration within the MLLM framework. GPT4Point as a powerful 3D MLLM seamlessly
can execute a variety of point-text reference tasks such as point-cloud captioni
ng and Q&A. Additionally GPT4Point is equipped with advanced capabilities for co
ntrollable 3D generation it can get high-quality results through a low-quality p
oint-text feature maintaining the geometric shapes and colors. To support the ex
pansive needs of 3D object-text pairs we develop Pyramid-XL a point-language dat
aset annotation engine. It constructs a large-scale database over 1M objects of
varied text granularity levels from the Objaverse-XL dataset essential for train
ing GPT4Point. A comprehensive benchmark has been proposed to evaluate 3D point-
language understanding capabilities. In extensive evaluations GPT4Point has demo
nstrated superior performance in understanding and generation.
****
SemCity: Semantic Scene Generation with Triplane Diffusion
Jumin Lee, Sebin Lee, Changho Jo, Woobin Im, Juhyeong Seon, Sung-Eui Yoon; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 28337-28347
We present "SemCity" a 3D diffusion model for semantic scene generation in real-
world outdoor environments. Most 3D diffusion models focus on generating a singl
e object synthetic indoor scenes or synthetic outdoor scenes while the generatio
n of real-world outdoor scenes is rarely addressed. In this paper we concentrate
on generating a real-outdoor scene through learning a diffusion model on a real
-world outdoor dataset. In contrast to synthetic data real-outdoor datasets ofte
n contain more empty spaces due to sensor limitations causing challenges in lear
ning real-outdoor distributions. To address this issue we exploit a triplane rep
resentation as a proxy form of scene distributions to be learned by our diffusio
n model. Furthermore we propose a triplane manipulation that integrates seamless
ly with our triplane diffusion model. The manipulation improves our diffusion mo
del's applicability in a variety of downstream tasks related to outdoor scene ge
neration such as scene inpainting scene outpainting and semantic scene completio
n refinements. In experimental results we demonstrate that our triplane diffusio
n model shows meaningful generation results compared with existing work in a rea
l-outdoor dataset SemanticKITTI. We also show our triplane manipulation facilita
tes seamlessly adding removing or modifying objects within a scene. Further it a
lso enables the expansion of scenes toward a city-level scale. Finally we evalua
te our method on semantic scene completion refinements where our diffusion model
enhances predictions of semantic scene completion networks by learning scene di
stribution. Our code is available at https://github.com/zoomin-lee/SemCity.
****
Improving Semantic Correspondence with Viewpoint-Guided Spherical Maps
Octave Mariotti, Oisin Mac Aodha, Hakan Bilen; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19521-19530
Recent self-supervised models produce visual features that are not only effectiv
e at encoding image-level but also pixel-level semantics. They have been reporte
d to obtain impressive results for dense visual semantic correspondence estimati
on even outperforming fully-supervised methods. Nevertheless these models still
fail in the presence of challenging image characteristics such as symmetries and
repeated parts. To address these limitations we propose a new semantic correspo
ndence estimation method that supplements state-of-the-art self-supervised featu
res with 3D understanding via a weak geometric spherical prior. Compared to more
involved 3D pipelines our model provides a simple and effective way of injectin
g informative geometric priors into the learned representation while requiring o
nly weak viewpoint information. We also propose a new evaluation metric that bet
ter accounts for repeated part and symmetry-induced mistakes. We show that our m
ethod succeeds in distinguishing between symmetric views and repeated parts acro
ss many object categories in the challenging SPair-71k dataset and also in gener
alizing to previously unseen classes in the AwA dataset.

****
MR-VNet: Media Restoration using Volterra Networks
Siddharth Roheda, Amit Unde, Loay Rashid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6098-6107
This research paper presents a novel class of restoration network architecture based on the Volterra series formulation. By incorporating non-linearity into the system response function through higher order convolutions instead of traditional activation functions we introduce a general framework for image/video restoration. Through extensive experimentation we demonstrate that our proposed architecture achieves state-of-the-art (SOTA) performance in the field of Image/Video Restoration. Moreover we establish that the recently introduced Non-Linear Activation Free Network (NAF-NET) can be considered a special case within the broader class of Volterra Neural Networks. These findings highlight the potential of Volterra Neural Networks as a versatile and powerful tool for addressing complex restoration tasks in computer vision.
****
Dual Memory Networks: A Versatile Adaptation Approach for Vision-Language Models
Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28718-28728
With the emergence of pre-trained vision-language models like CLIP how to adapt them to various downstream classification tasks has garnered significant attention in recent research. The adaptation strategies can be typically categorized into three paradigms: zero-shot adaptation few-shot adaptation and the recently-proposed training-free few-shot adaptation. Most existing approaches are tailored for a specific setting and can only cater to one or two of these paradigms. In this paper we introduce a versatile adaptation approach that can effectively work under all three settings. Specifically we propose the dual memory networks that comprise dynamic and static memory components. The static memory caches training data knowledge enabling training-free few-shot adaptation while the dynamic memory preserves historical test features online during the testing process allowing for the exploration of additional data insights beyond the training set. This novel capability enhances model performance in the few-shot setting and enables model usability in the absence of training data. The two memory networks employ the same flexible memory interactive strategy which can operate in a training-free mode and can be further enhanced by incorporating learnable projection layers. Our approach is tested across 11 datasets under the three task settings. Remarkably in the zero-shot scenario it outperforms existing methods by over 3% and even shows superior results against methods utilizing external training data. Additionally our method exhibits robust performance against natural distribution shifts.
****
Single Mesh Diffusion Models with Field Latents for Texture Generation
Thomas W. Mitchel, Carlos Esteves, Ameesh Makadia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7953-7963
We introduce a framework for intrinsic latent diffusion models operating directly on the surfaces of 3D shapes with the goal of synthesizing high-quality textures. Our approach is underpinned by two contributions: Field Latents a latent representation encoding textures as discrete vector fields on the mesh vertices and Field Latent Diffusion Models which learn to denoise a diffusion process in the learned latent space on the surface. We consider a single-textured-mesh paradigm where our models are trained to generate variations of a given texture on a mesh. We show the synthesized textures are of superior fidelity compared those from existing single-textured-mesh generative models. Our models can also be adapted for user-controlled editing tasks such as inpainting and label-guided generation. The efficacy of our approach is due in part to the equivariance of our proposed framework under isometries allowing our models to seamlessly reproduce details across locally similar regions and opening the door to a notion of generative texture transfer. Code and visualizations are available at https://single-mesh-diffusion.github.io/.

****
LION: Empowering Multimodal Large Language Model with Dual-Level Visual Knowledg
e
Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, Liqiang Nie; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 26540-26550
Multimodal Large Language Models (MLLMs) have endowed LLMs with the ability to p
erceive and understand multi-modal signals. However most of the existing MLLMs m
ainly adopt vision encoders pretrained on coarsely aligned image-text pairs lead
ing to insufficient extraction and reasoning of visual knowledge. To address thi
s issue we devise a dual-Level vIsual knOwledge eNhanced Multimodal Large Langua
ge Model (LION) which empowers the MLLM by injecting visual knowledge in two lev
els. 1) Progressive incorporation of fine-grained spatial-aware visual knowledge
. We design a vision aggregator cooperated with region-level vision-language (VL
) tasks to incorporate fine-grained spatial-aware visual knowledge into the MLLM
. To alleviate the conflict between image-level and region-level VL tasks during
incorporation we devise a dedicated stage-wise instruction-tuning strategy with
mixture-of-adapters. This progressive incorporation scheme contributes to the m
utual promotion between these two kinds of VL tasks. 2) Soft prompting of high-l
evel semantic visual evidence. We facilitate the MLLM with high-level semantic v
isual evidence by leveraging diverse image tags. To mitigate the potential influ
ence caused by imperfect predicted tags we propose a soft prompting method by em
bedding a learnable token into the tailored text instruction. Comprehensive expe
riments on several multi-modal benchmarks demonstrate the superiority of our mod
el (e.g. improvement of 5% accuracy on VSR and 3% CIDEr on TextCaps over Instruc
tBLIP 5% accuracy on RefCOCOg over Kosmos-2).
****
Learning to Select Views for Efficient Multi-View Understanding
Yunzhong Hou, Stephen Gould, Liang Zheng; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20135-20144
Multiple camera view (multi-view) setups have proven useful in many computer vis
ion applications. However the high computational cost associated with multiple v
iews creates a significant challenge for end devices with limited computational
resources. In modern CPU pipelining breaks a longer job into steps and enables p
arallelism over sequential steps from multiple jobs. Inspired by this we study s
elective view pipelining for efficient multi-view understanding which breaks com
putation of multiple views into steps and only computes the most helpful views/s
teps in a parallel manner for the best efficiency. To this end we use reinforcem
ent learning to learn a very light view selection module that analyzes the targe
t object or scenario from initial views and selects the next-best-view for recog
nition or detection for pipeline computation. Experimental results on multi-view
classification and detection tasks show that our approach achieves promising pe
rformance while using only 2 or 3 out of N available views significantly reducin
g computational costs while maintaining parallelism over GPU through selective v
iew pipelining.
****
Consistency and Uncertainty: Identifying Unreliable Responses From Black-Box Vis
ion-Language Models for Selective Visual Question Answering
Zaid Khan, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2024, pp. 10854-10863
The goal of selective prediction is to allow an a model to abstain when it may n
ot be able to deliver a reliable prediction which is important in safety-critica
l contexts. Existing approaches to selective prediction typically require access
to the internals of a model require retraining a model or study only unimodal m
odels. However the most powerful models (e.g. GPT-4) are typically only availabl
e as black boxes with inaccessible internals are not retrainable by end-users an
d are frequently used for multimodal tasks. We study the possibility of selectiv
e prediction for vision-language models in a realistic black-box setting. We pro
pose using the principle of neighborhood consistency to identify unreliable resp
onses from a black-box vision-language model in question answering tasks. We hyp

othesize that given only a visual question and model response the consistency of the model's responses over the neighborhood of a visual question will indicate reliability. It is impossible to directly sample neighbors in feature space in a black-box setting. Instead we show that it is possible to use a smaller proxy model to approximately sample from the neighborhood. We find that neighborhood consistency can be used to identify model responses to visual questions that are likely unreliable even in adversarial settings or settings that are out-of-distribution to the proxy model.
****
SAI3D: Segment Any Instance in 3D Scenes
Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, Baoquan Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3292-3302
Advancements in 3D instance segmentation have traditionally been tethered to the availability of annotated datasets limiting their application to a narrow spectrum of object categories. Recent efforts have sought to harness vision-language models like CLIP for open-set semantic reasoning yet these methods struggle to distinguish between objects of the same categories and rely on specific prompts that are not universally applicable. In this paper we introduce SAI3D a novel zero-shot 3D instance segmentation approach that synergistically leverages geometric priors and semantic cues derived from Segment Anything Model (SAM). Our method partitions a 3D scene into geometric primitives which are then progressively merged into 3D instance segmentations that are consistent with the multi-view SAM masks. Moreover we design a hierarchical region-growing algorithm with a dynamic thresholding mechanism which largely improves the robustness of fine-grained 3D scene parsing. Empirical evaluations on ScanNet Matterport3D and the more challenging ScanNet++ datasets demonstrate the superiority of our approach. Notably SAI3D outperforms existing open-vocabulary baselines and even surpasses fully-supervised methods in class-agnostic segmentation on ScanNet++. Our project page is at https://yd-yin.github.io/SAI3D/.
****
Implicit Motion Function
Yue Gao, Jiahao Li, Lei Chu, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19278-19289
Recent advancements in video modeling extensively rely on optical flow to represent the relationships across frames but this approach often lacks efficiency and fails to model the probability of the intrinsic motion of objects. In addition conventional encoder-decoder frameworks in video processing focus on modeling the correlation in the encoder leading to limited generative capabilities and redundant intermediate representations. To address these challenges this paper proposes a novel Implicit Motion Function (IMF) method. Our approach utilizes a low-dimensional latent token as the implicit representation along with the use of cross-attention to implicitly model the correlation between frames. This enables the implicit modeling of temporal correlations and understanding of object motions. Our method not only improves sparsity and efficiency in representation but also explores the generative capabilities of the decoder by integrating correlation modeling within it. The IMF framework facilitates video editing and other generative tasks by allowing the direct manipulation of latent tokens. We validate the effectiveness of IMF through extensive experiments on multiple video tasks demonstrating superior performance in terms of reconstructed video quality compression efficiency and generation ability.
****
Unified Entropy Optimization for Open-Set Test-Time Adaptation
Zhengqing Gao, Xu-Yao Zhang, Cheng-Lin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23975-23984
Test-time adaptation (TTA) aims at adapting a model pre-trained on the labeled source domain to the unlabeled target domain. Existing methods usually focus on improving TTA performance under covariate shifts while neglecting semantic shifts. In this paper we delve into a realistic open-set TTA setting where the target domain may contain samples from unknown classes. Many state-of-the-art closed-se

t TTA methods perform poorly when applied to open-set scenarios which can be att
ributed to the inaccurate estimation of data distribution and model confidence.
To address these issues we propose a simple but effective framework called unifi
ed entropy optimization (UniEnt) which is capable of simultaneously adapting to
covariate-shifted in-distribution (csID) data and detecting covariate-shifted ou
t-of-distribution (csOOD) data. Specifically UniEnt first mines pseudo-csID and
pseudo-csOOD samples from test data followed by entropy minimization on the pseu
do-csID data and entropy maximization on the pseudo-csOOD data. Furthermore we i
ntroduce UniEnt+ to alleviate the noise caused by hard data partition leveraging
 sample-level confidence. Extensive experiments on CIFAR benchmarks and Tiny-Ima
geNet-C show the superiority of our framework. The code is available at https://
github.com/gaozhengqing/UniEnt.
****
TexOct: Generating Textures of 3D Models with Octree-based Diffusion
Jialun Liu, Chenming Wu, Xinqi Liu, Xing Liu, Jinbo Wu, Haotian Peng, Chen Zhao,
 Haocheng Feng, Jingtuo Liu, Errui Ding; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4284-4293
This paper focuses on synthesizing high-quality and complete textures directly o
n the surface of 3D models within 3D space. 2D diffusion-based methods face chal
lenges in generating 2D texture maps due to the infinite possibilities of UV map
ping for a given 3D mesh. Utilizing point clouds helps circumvent variations ari
sing from diverse mesh topologies and UV mappings. Nevertheless achieving dense
point clouds to accurately represent texture details poses a challenge due to li
mited computational resources. To address these challenges we propose an efficie
nt octree-based diffusion pipeline called TexOct. Our method starts by sampling
a point cloud from the surface of a given 3D model with each point containing te
xture noise values. We utilize an octree structure to efficiently represent this
 point cloud. Additionally we introduce an innovative octree-based diffusion mod
el that leverages the denoising capabilities of the Denoising Diffusion Probabil
istic Model (DDPM). This model gradually reduces the texture noise on the octree
 nodes resulting in the restoration of fine texture. Experimental results on Sha
peNet demonstrate that TexOct effectively generates high-quality 3D textures in
both unconditional and text / image-conditional scenarios.
****
Anatomically Constrained Implicit Face Models
Prashanth Chandran, Gaspard Zoss; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2024, pp. 2220-2229
Coordinate based implicit neural representations have gained rapid popularity in
 recent years as they have been successfully used in image geometry and scene mo
deling tasks. In this work we present a novel use case for such implicit represe
ntations in the context of learning anatomically constrained face models. Actor
specific anatomically constrained face models are the state of the art in both f
acial performance capture and performance retargeting. Despite their practical s
uccess these anatomical models are slow to evaluate and often require extensive
data capture to be built. We propose the anatomical implicit face model; an ense
mble of implicit neural networks that jointly learn to model the facial anatomy
and the skin surface with high-fidelity and can readily be used as a drop in rep
lacement to conventional blendshape models. Given an arbitrary set of skin surfa
ce meshes of an actor and only a neutral shape with estimated skull and jaw bone
s our method can recover a dense anatomical substructure which constrains every
point on the facial surface. We demonstrate the usefulness of our approach in se
veral tasks ranging from shape fitting shape editing and performance retargeting
.
****
Expandable Subspace Ensemble for Pre-Trained Model-Based Class-Incremental Learn
ing
Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, De-Chuan Zhan; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23554-
23564
Class-Incremental Learning (CIL) requires a learning system to continually learn

new classes without forgetting. Despite the strong performance of Pre-Trained Models (PTMs) in CIL a critical issue persists: learning new classes often results in the overwriting of old ones. Excessive modification of the network causes forgetting while minimal adjustments lead to an inadequate fit for new classes. As a result it is desired to figure out a way of efficient model updating without harming former knowledge. In this paper we propose ExpAndable Subspace Ensemble (EASE) for PTM-based CIL. To enable model updating without conflict we train a distinct lightweight adapter module for each new task aiming to create task-specific subspaces. These adapters span a high-dimensional feature space enabling joint decision-making across multiple subspaces. As data evolves the expanding subspaces render the old class classifiers incompatible with new-stage spaces. Correspondingly we design a semantic-guided prototype complement strategy that synthesizes old classes' new features without using any old class instance. Extensive experiments on seven benchmark datasets verify EASE's state-of-the-art performance. Code is available at: https://github.com/sun-hailong/CVPR24-Ease
****
Capturing Closely Interacted Two-Person Motions with Reaction Priors
Qi Fang, Yinghui Fan, Yanjun Li, Junting Dong, Dingwei Wu, Weidong Zhang, Kang Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 655-665
In this paper we focus on capturing closely interacted two-person motions from monocular videos an important yet understudied topic. Unlike less-interacted motions closely interacted motions contain frequently occurring inter-human occlusions which pose significant challenges to existing capturing algorithms. To address this problem our key observation is that close physical interactions between two subjects typically happen under very specific situations (e.g. handshake hug etc.) and such situational contexts contain strong prior semantics to help infer the poses of occluded joints. In this spirit we introduce reaction priors which are invertible neural networks that bi-directionally model the pose probability distributions of one person given the pose of the other. The learned reaction priors are then incorporated into a query-based pose estimator which is a decoder-only Transformer with self-attentions on both intra-joint and inter-joint relationships. We demonstrate that our design achieves considerably higher performance than previous methods on multiple benchmarks. What's more as existing datasets lack sufficient cases of close human-human interactions we also build a new dataset called Dual-Human to better evaluate different methods. Dual-Human contains around 2k sequences of closely interacted two-person motions each with synthetic multi-view renderings contact annotations and text descriptions. We believe that this new public dataset can significantly promote further research in this area.
****
RobustSAM: Segment Anything Robustly on Degraded Images
Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhou Ma, Jian Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4081-4091
Segment Anything Model (SAM) has emerged as a transformative approach in image segmentation acclaimed for its robust zero-shot segmentation capabilities and flexible prompting system. Nonetheless its performance is challenged by images with degraded quality. Addressing this limitation we propose the Robust Segment Anything Model (RobustSAM) which enhances SAM's performance on low-quality images while preserving its promptability and zero-shot generalization. Our method leverages the pre-trained SAM model with only marginal parameter increments and computational requirements. The additional parameters of RobustSAM can be optimized within 30 hours on eight GPUs demonstrating its feasibility and practicality for typical research laboratories. We also introduce the Robust-Seg dataset a collection of 688K image-mask pairs with different degradations designed to train and evaluate our model optimally. Extensive experiments across various segmentation tasks and datasets confirm RobustSAM's superior performance especially under zero-shot conditions underscoring its potential for extensive real-world application. Additionally our method has been shown to effectively improve the performance

of SAM-based downstream tasks such as single image dehazing and deblurring.
****
MultiDiff: Consistent Novel View Synthesis from a Single Image
Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, Peter Kontschieder; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10258-10268
We introduce MultiDiff a novel approach for consistent novel view synthesis of scenes from a single RGB image. The task of synthesizing novel views from a single reference image is highly ill-posed by nature as there exist multiple plausible explanations for unobserved areas. To address this issue we incorporate strong priors in form of monocular depth predictors and video-diffusion models. Monocular depth enables us to condition our model on warped reference images for the target views increasing geometric stability. The video-diffusion prior provides a strong proxy for 3D scenes allowing the model to learn continuous and pixel-accurate correspondences across generated images. In contrast to approaches relying on autoregressive image generation that are prone to drifts and error accumulation MultiDiff jointly synthesizes a sequence of frames yielding high-quality and multi-view consistent results -- even for long-term scene generation with large camera movements while reducing inference time by an order of magnitude. For additional consistency and image quality improvements we introduce a novel structured noise distribution. Our experimental results demonstrate that MultiDiff outperforms state-of-the-art methods on the challenging real-world datasets RealEstate10K and ScanNet. Finally our model naturally supports multi-view consistent editing without the need for further tuning.
****
In-N-Out: Faithful 3D GAN Inversion with Volumetric Decomposition for Face Editing
Yiran Xu, Zhixin Shu, Cameron Smith, Seoung Wug Oh, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7225-7235
3D-aware GANs offer new capabilities for view synthesis while preserving the editing functionalities of their 2D counterparts. GAN inversion is a crucial step that seeks the latent code to reconstruct input images or videos subsequently enabling diverse editing tasks through manipulation of this latent code. However a model pre-trained on a particular dataset (e.g. FFHQ) often has difficulty reconstructing images with out-of-distribution (OOD) objects such as faces with heavy make-up or occluding objects. We address this issue by explicitly modeling OOD objects from the input in 3D-aware GANs. Our core idea is to represent the image using two individual neural radiance fields: one for the in-distribution content and the other for the out-of-distribution object. The final reconstruction is achieved by optimizing the composition of these two radiance fields with carefully designed regularization. We demonstrate that our explicit decomposition alleviates the inherent trade-off between reconstruction fidelity and editability. We evaluate reconstruction accuracy and editability of our method on challenging real face images and videos and showcase favorable results against other baselines.
****
Atom-Level Optical Chemical Structure Recognition with Limited Supervision
Martijn Oldenhof, Edward De Brouwer, Adam Arany, Yves Moreau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17669-17678
Identifying the chemical structure from a graphical representation or image of a molecule is a challenging pattern recognition task that would greatly benefit drug development. Yet existing methods for chemical structure recognition do not typically generalize well and show diminished effectiveness when confronted with domains where data is sparse or costly to generate such as hand-drawn molecule images. To address this limitation we propose a new chemical structure recognition tool that delivers state-of-the-art performance and can adapt to new domains with a limited number of data samples and supervision. Unlike previous approaches our method provides atom-level localization and can therefore segment the imag

e into the different atoms and bonds. Our model is the first model to perform OCSR with atom-level entity detection with only SMILES supervision. Through rigorous and extensive benchmarking we demonstrate the preeminence of our chemical structure recognition approach in terms of data efficiency accuracy and atom-level entity prediction.
****
L4D-Track: Language-to-4D Modeling Towards 6-DoF Tracking and Shape Reconstruction in 3D Point Cloud Stream
Jingtao Sun, Yaonan Wang, Mingtao Feng, Yulan Guo, Ajmal Mian, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21146-21156
3D visual language multi-modal modeling plays an important role in actual human-computer interaction. However the inaccessibility of large-scale 3D-language pairs restricts their applicability in real-world scenarios. In this paper we aim to handle a real-time multi-task for 6-DoF pose tracking of unknown objects leveraging 3D-language pre-training scheme from a series of 3D point cloud video streams while simultaneously performing 3D shape reconstruction in current observation. To this end we present a generic Language-to-4D modeling paradigm termed L4D-Track that tackles zero-shot 6-DoF \underline Track ing and shape reconstruction by learning pairwise implicit 3D representation and multi-level multi-modal alignment. Our method constitutes two core parts. 1) Pairwise Implicit 3D Space Representation that establishes spatial-temporal to language coherence descriptions across continuous 3D point cloud video. 2) Language-to-4D Association and Contrastive Alignment enables multi-modality semantic connections between 3D point cloud video and language. Our method trained exclusively on public NOCS-REAL275 dataset achieves promising results on both two publicly benchmarks. This not only shows powerful generalization performance but also proves its remarkable capability in zero-shot inference.
****
General Point Model Pretraining with Autoencoding and Autoregressive
Zhe Li, Zhangyang Gao, Cheng Tan, Bocheng Ren, Laurence T. Yang, Stan Z. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20954-20964
The pre-training architectures of large language models encompass various types including autoencoding models autoregressive models and encoder-decoder models. We posit that any modality can potentially benefit from a large language model as long as it undergoes vector quantization to become discrete tokens. Inspired by the General Language Model we propose a General Point Model (GPM) that seamlessly integrates autoencoding and autoregressive tasks in a point cloud transformer. This model is versatile allowing fine-tuning for downstream point cloud representation tasks as well as unconditional and conditional generation tasks. GPM enhances masked prediction in autoencoding through various forms of mask padding tasks leading to improved performance in point cloud understanding. Additionally GPM demonstrates highly competitive results in unconditional point cloud generation tasks even exhibiting the potential for conditional generation tasks by modifying the input's conditional information. Compared to models like Point-BERT MaskPoint and PointMAE our GPM achieves superior performance in point cloud understanding tasks. Furthermore the integration of autoregressive and autoencoding within the same transformer underscores its versatility across different downstream tasks.
****
Combining Frame and GOP Embeddings for Neural Video Representation
Jens Eirik Saethre, Roberto Azevedo, Christopher Schroers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9253-9263
Implicit neural representations (INRs) were recently proposed as a new video compression paradigm with existing approaches performing on par with HEVC. However such methods only perform well in limited settings e.g. specific model sizes fixed aspect ratios and low-motion videos. We address this issue by proposing T-NeRV a hybrid video INR that combines frame-specific embeddings with GOP-specific f

eatures providing a lever for content-specific fine-tuning. We employ entropy-co nstrained training to jointly optimize our model for rate and distortion and dem onstrate that T-NeRV can thereby automatically adjust this lever during training effectively fine-tuning itself to the target content. We evaluate T-NeRV on the UVG dataset where it achieves state-of-the-art results on the video representat ion task outperforming previous works by up to 3dB PSNR on challenging high-moti on sequences. Further our method improves on the compression performance of prev ious methods and is the first video INR to outperform HEVC on all UVG sequences.
****

LiDAR-based Person Re-identification
Wenxuan Guo, Zhiyu Pan, Yingping Liang, Ziheng Xi, Zhicheng Zhong, Jianjiang Fen g, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2024, pp. 17437-17447
Camera-based person re-identification (ReID) systems have been widely applied in the field of public security. However cameras often lack the perception of 3D m orphological information of human and are susceptible to various limitations suc h as inadequate illumination complex background and personal privacy. In this pa per we propose a LiDAR-based ReID framework ReID3D that utilizes pre-training st rategy to retrieve features of 3D body shape and introduces Graph-based Compleme ntary Enhancement Encoder for extracting comprehensive features. Due to the lack of LiDAR datasets we build LReID the first LiDAR-based person ReID dataset whic h is collected in several outdoor scenes with variations in natural conditions. Additionally we introduce LReID-sync a simulated pedestrian dataset designed for pre-training encoders with tasks of point cloud completion and shape parameter learning. Extensive experiments on LReID show that ReID3D achieves exceptional p erformance with a rank-1 accuracy of 94.0 highlighting the significant potential of LiDAR in addressing person ReID tasks. To the best of our knowledge we are t he first to propose a solution for LiDAR-based ReID. The code and dataset are av ailable at https://github.com/GWxuan/ReID3D.
****

Fantastic Animals and Where to Find Them: Segment Any Marine Animal with Dual SA M
Pingping Zhang, Tianyu Yan, Yang Liu, Huchuan Lu; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2578-2587
As an important pillar of underwater intelligence Marine Animal Segmentation (MA S) involves segmenting animals within marine environments. Previous methods don' t excel in extracting long-range contextual features and overlook the connectivi ty between discrete pixels. Recently Segment Anything Model (SAM) offers a unive rsal framework for general segmentation tasks. Unfortunately trained with natura l images SAM does not obtain the prior knowledge from marine images. In addition the single-position prompt of SAM is very insufficient for prior guidance. To a ddress these issues we propose a novel feature learning framework named Dual-SAM for high-performance MAS. To this end we first introduce a dual structure with SAM's paradigm to enhance feature learning of marine images. Then we propose a M ulti-level Coupled Prompt (MCP) strategy to instruct comprehensive underwater pr ior information and enhance the multi-level features of SAM's encoder with adapt ers. Subsequently we design a Dilated Fusion Attention Module (DFAM) to progress ively integrate multi-level features from SAM's encoder. Finally instead of dire ctly predicting the masks of marine animals we propose a Criss-Cross Connectivit y Prediction (C3P) paradigm to capture the inter-connectivity between discrete p ixels. With dual decoders it generates pseudo-labels and achieves mutual supervi sion for complementary feature representations resulting in considerable improve ments over previous techniques. Extensive experiments verify that our proposed m ethod achieves state-of-the-art performances on five widely-used MAS datasets. T he code is available at https://github.com/Drchip61/Dual SAM.
****

Seeing and Hearing: Open-domain Visual-Audio Generation with Diffusion Latent Al igners
Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, Qifeng Chen; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,

pp. 7151-7161
Video and audio content creation serves as the core technique for the movie industry and professional users. Recently existing diffusion-based methods tackle video and audio generation separately which hinders the technique transfer from academia to industry. In this work we aim at filling the gap with a carefully designed optimization-based framework for cross-visual-audio and joint-visual-audio generation. We observe the powerful generation ability of off-the-shelf video or audio generation models. Thus instead of training the giant models from scratch we propose to bridge the existing strong models with a shared latent representation space. Specifically we propose a multimodality latent aligner with the pre-trained ImageBind model. Our latent aligner shares a similar core as the classifier guidance that guides the diffusion denoising process during inference time. Through carefully designed optimization strategy and loss functions we show the superior performance of our method on joint video-audio generation visual-steered audio generation and audio-steered visual generation tasks. The project website can be found at \href https://yzxing87.github.io/Seeing-and-Hearing/ https://yzxing87.github.io/Seeing-and-Hearing/ .
****
Model Adaptation for Time Constrained Embodied Control
Jaehyun Song, Minjong Yoo, Honguk Woo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16499-16508
When adopting a deep learning model for embodied agents it is required that the model structure be optimized for specific tasks and operational conditions. Such optimization can be static such as model compression or dynamic such as adaptive inference. Yet these techniques have not been fully investigated for embodied control systems subject to time constraints which necessitate sequential decision-making for multiple tasks each with distinct inference latency limitations. In this paper we present MoDeC a time constraint-aware embodied control framework using the modular model adaptation. We formulate model adaptation to varying operational conditions on resource and time restrictions as dynamic routing on a modular network incorporating these conditions as part of multi-task objectives. Our evaluation across several vision-based embodied environments demonstrates the robustness of MoDeC showing that it outperforms other model adaptation methods in both performance and adherence to time constraints in robotic manipulation and autonomous driving applications.
****
Objects as Volumes: A Stochastic Geometry View of Opaque Solids
Bailey Miller, Hanyu Chen, Alice Lai, Ioannis Gkioulekas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 87-97
We develop a theory for the representation of opaque solids as volumes. Starting from a stochastic representation of opaque solids as random indicator functions we prove the conditions under which such solids can be modeled using exponential volumetric transport. We also derive expressions for the volumetric attenuation coefficient as a functional of the probability distributions of the underlying indicator functions. We generalize our theory to account for isotropic and anisotropic scattering at different parts of the solid and for representations of opaque solids as stochastic implicit surfaces. We derive our volumetric representation from first principles which ensures that it satisfies physical constraints such as reciprocity and reversibility. We use our theory to explain compare and correct previous volumetric representations as well as propose meaningful extensions that lead to improved performance in 3D reconstruction tasks.
****
ActiveDC: Distribution Calibration for Active Finetuning
Wenshuai Xu, Zhenghui Hu, Yu Lu, Jinzhou Meng, Qingjie Liu, Yunhong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16996-17005
The pretraining-finetuning paradigm has gained popularity in various computer vision tasks. In this paradigm the emergence of active finetuning arises due to the abundance of large-scale data and costly annotation requirements. Active finet

uning involves selecting a subset of data from an unlabeled pool for annotation facilitating subsequent finetuning. However the use of a limited number of training samples can lead to a biased distribution potentially resulting in model overfitting. In this paper we propose a new method called ActiveDC for the active finetuning tasks. Firstly we select samples for annotation by optimizing the distribution similarity between the subset to be selected and the entire unlabeled pool in continuous space. Secondly we calibrate the distribution of the selected samples by exploiting implicit category information in the unlabeled pool. The feature visualization provides an intuitive sense of the effectiveness of our approach to distribution calibration. We conducted extensive experiments on three image classification datasets with different sampling ratios. The results indicate that ActiveDC consistently outperforms the baseline performance in all image classification tasks. The improvement is particularly significant when the sampling ratio is low with performance gains of up to 10%. Our code will be released.
****
Seeing Unseen: Discover Novel Biomedical Concepts via Geometry-Constrained Probabilistic Modeling
Jianan Fan, Dongnan Liu, Hang Chang, Heng Huang, Mei Chen, Weidong Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11524-11534
Machine learning holds tremendous promise for transforming the fundamental practice of scientific discovery by virtue of its data-driven nature. With the ever-increasing stream of research data collection it would be appealing to autonomously explore patterns and insights from observational data for discovering novel classes of phenotypes and concepts. However in the biomedical domain there are several challenges inherently presented in the cumulated data which hamper the progress of novel class discovery. The non-i.i.d. data distribution accompanied by the severe imbalance among different groups of classes essentially leads to ambiguous and biased semantic representations. In this work we present a geometry-constrained probabilistic modeling treatment to resolve the identified issues. First we propose to parameterize the approximated posterior of instance embedding as a marginal von Mises-Fisher distribution to account for the interference of distributional latent bias. Then we incorporate a suite of critical geometric properties to impose proper constraints on the layout of constructed embedding space which in turn minimizes the uncontrollable risk for unknown class learning and structuring. Furthermore a spectral graph-theoretic method is devised to estimate the number of potential novel classes. It inherits two intriguing merits compared to existent approaches namely high computational efficiency and flexibility for taxonomy-adaptive estimation. Extensive experiments across various biomedical scenarios substantiate the effectiveness and general applicability of our method.
****
MVHumanNet: A Large-scale Dataset of Multi-view Daily Dressing Human Captures
Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, Shuguang Cui, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19801-19811
In this era the success of large language models and text-to-image models can be attributed to the driving force of large-scale datasets. However in the realm of 3D vision while remarkable progress has been made with models trained on large-scale synthetic and real-captured object data like Objaverse and MVImgNet a similar level of progress has not been observed in the domain of human-centric tasks partially due to the lack of a large-scale human dataset. Existing datasets of high-fidelity 3D human capture continue to be mid-sized due to the significant challenges in acquiring large-scale high-quality 3D human data. To bridge this gap we present MVHumanNet a dataset that comprises multi-view human action sequences of 4500 human identities. The primary focus of our work is on collecting human data that features a large number of diverse identities and everyday clothing using a multi-view human capture system which facilitates easily scalable data collection. Our dataset contains 9000 daily outfits 60000 motion sequences and 6

45 million frames with extensive annotations including human masks camera parameters 2D and 3D keypoints SMPL/SMPLX parameters and corresponding textual descriptions. To explore the potential of MVHumanNet in various 2D and 3D visual tasks we conducted pilot studies on view-consistent action recognition human NeRF reconstruction text-driven view-unconstrained human image generation as well as 2D view-unconstrained human image and 3D avatar generation. Extensive experiments demonstrate the performance improvements and effective applications enabled by the scale provided by MVHumanNet. As the current largest-scale 3D human dataset we hope that the release of MVHumanNet data with annotations will foster further innovations in the domain of 3D human-centric tasks at scale.
****
Communication-Efficient Federated Learning with Accelerated Client Gradient
Geeho Kim, Jinkyu Kim, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12385-12394
Federated learning often suffers from slow and unstable convergence due to the heterogeneous characteristics of participating client datasets. Such a tendency is aggravated when the client participation ratio is low since the information collected from the clients has large variations. To address this challenge we propose a simple but effective federated learning framework which improves the consistency across clients and facilitates the convergence of the server model. This is achieved by making the server broadcast a global model with a lookahead gradient. This strategy enables the proposed approach to convey the projected global update information to participants effectively without additional client memory and extra communication costs. We also regularize local updates by aligning each client with the overshot global model to reduce bias and improve the stability of our algorithm. We provide the theoretical convergence rate of our algorithm and demonstrate remarkable performance gains in terms of accuracy and communication efficiency compared to the state-of-the-art methods especially with low client participation rates. The source code is available at our project page.
****
LLMs are Good Action Recognizers
Haoxuan Qu, Yujun Cai, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18395-18406
Skeleton-based action recognition has attracted lots of research attention. Recently to build an accurate skeleton-based action recognizer a variety of works have been proposed. Among them some works use large model architectures as backbones of their recognizers to boost the skeleton data representation capability while some other works pre-train their recognizers on external data to enrich the knowledge. In this work we observe that large language models which have been extensively used in various natural language processing tasks generally hold both large model architectures and rich implicit knowledge. Motivated by this we propose a novel LLM-AR framework in which we investigate treating the Large Language Model as an Action Recognizer. In our framework we propose a linguistic projection process to project each input action signal (i.e. each skeleton sequence) into its "sentence format" (i.e. an "action sentence"). Moreover we also incorporate our framework with several designs to further facilitate this linguistic projection process. Extensive experiments demonstrate the efficacy of our proposed framework.
****
NoiseCLR: A Contrastive Learning Approach for Unsupervised Discovery of Interpretable Directions in Diffusion Models
Yusuf Dalva, Pinar Yanardag; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24209-24218
Generative models have been very popular in the recent years for their image generation capabilities. GAN-based models are highly regarded for their disentangled latent space which is a key feature contributing to their success in controlled image editing. On the other hand diffusion models have emerged as powerful tools for generating high-quality images. However the latent space of diffusion models is not as thoroughly explored or understood. Existing methods that aim to explore the latent space of diffusion models usually relies on text prompts to pin

point specific semantics. However this approach may be restrictive in areas such as art fashion or specialized fields like medicine where suitable text prompts might not be available or easy to conceive thus limiting the scope of existing work. In this paper we propose an unsupervised method to discover latent semantics in text-to-image diffusion models without relying on text prompts. Our method takes a small set of unlabeled images from specific domains such as faces or cats and a pre-trained diffusion model and discovers diverse semantics in unsupervised fashion using a contrastive learning objective. Moreover the learned directions can be applied simultaneously either within the same domain (such as various types of facial edits) or across different domains (such as applying cat and face edits within the same image) without interfering with each other. Our extensive experiments show that our method achieves highly disentangled edits outperforming existing approaches in both diffusion-based and GAN-based latent space editing methods.
****

SpecNeRF: Gaussian Directional Encoding for Specular Reflections
Li Ma, Vasu Agrawal, Haithem Turki, Changil Kim, Chen Gao, Pedro Sander, Michael Zollhöfer, Christian Richardt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21188-21198
Neural radiance fields have achieved remarkable performance in modeling the appearance of 3D scenes. However existing approaches still struggle with the view-dependent appearance of glossy surfaces especially under complex lighting of indoor environments. Unlike existing methods which typically assume distant lighting like an environment map we propose a learnable Gaussian directional encoding to better model the view-dependent effects under near-field lighting conditions. Importantly our new directional encoding captures the spatially-varying nature of near-field lighting and emulates the behavior of prefiltered environment maps. As a result it enables the efficient evaluation of preconvolved specular color at any 3D location with varying roughness coefficients. We further introduce a data-driven geometry prior that helps alleviate the shape radiance ambiguity in reflection modeling. We show that our Gaussian directional encoding and geometry prior significantly improve the modeling of challenging specular reflections in neural radiance fields which helps decompose appearance into more physically meaningful components.
****

Improving Subject-Driven Image Synthesis with Subject-Agnostic Guidance
Kelvin C.K. Chan, Yang Zhao, Xuhui Jia, Ming-Hsuan Yang, Huisheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6733-6742
In subject-driven text-to-image synthesis the synthesis process tends to be heavily influenced by the reference images provided by users often overlooking crucial attributes detailed in the text prompt. In this work we propose Subject-Agnostic Guidance (SAG) a simple yet effective solution to remedy the problem. We show that through constructing a subject-agnostic condition and applying our proposed dual classifier-free guidance one could obtain outputs consistent with both the given subject and input text prompts. We validate the efficacy of our approach through both optimization-based and encoder-based methods. Additionally we demonstrate its applicability in second-order customization methods where an encoder-based model is fine-tuned with DreamBooth. Our approach is conceptually simple and requires only minimal code modifications but leads to substantial quality improvements as evidenced by our evaluations and user studies.
****

Diffusion Model Alignment Using Direct Preference Optimization
Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, Nikhil Naik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8228-8238
Large language models (LLMs) are fine-tuned using human comparison data with Reinforcement Learning from Human Feedback (RLHF) methods to make them better aligned with users' preferences. In contrast to LLMs human preference learning has no

t been widely explored in text-to-image diffusion models; the best existing approach is to fine-tune a pretrained model using carefully curated high quality images and captions to improve visual appeal and text alignment. We propose Diffusion-DPO a method to align diffusion models to human preferences by directly optimizing on human comparison data. Diffusion-DPO is adapted from the recently developed Direct Preference Optimization (DPO) a simpler alternative to RLHF which directly optimizes a policy that best satisfies human preferences under a classification objective. We re-formulate DPO to account for a diffusion model notion of likelihood utilizing the evidence lower bound to derive a differentiable objective. Using the Pick-a-pic dataset of 851K crowdsourced pairwise preferences we fine-tune the base model of the state-of-the-art Stable Diffusion XL (SDXL)-1.0 model with Diffusion-DPO. Our fine-tuned base model significantly outperforms both base SDXL-1.0 and the larger SDXL-1.0 model consisting of an additional refinement model in human evaluation improving visual appeal and prompt alignment. We also develop a variant that uses AI feedback and has comparable performance to training on human preferences opening the door for scaling of diffusion model alignment methods.
****

Interactive Continual Learning: Fast and Slow Thinking

Biqing Qi, Xinquan Chen, Junqi Gao, Dong Li, Jianxing Liu, Ligang Wu, Bowen Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12882-12892

Advanced life forms sustained by the synergistic interaction of neural cognitive mechanisms continually acquire and transfer knowledge throughout their lifespan. In contrast contemporary machine learning paradigms exhibit limitations in emulating the facets of continual learning (CL). Nonetheless the emergence of large language models (LLMs) presents promising avenues for realizing CL via interactions with these models. Drawing on Complementary Learning System theory this paper presents a novel Interactive Continual Learning (ICL) framework enabled by collaborative interactions among models of various sizes. Specifically we assign the ViT model as System1 and multimodal LLM as System2. To enable the memory module to deduce tasks from class information and enhance Set2Set retrieval we propose the Class-Knowledge-Task Multi-Head Attention (CKT-MHA). Additionally to improve memory retrieval in System1 through enhanced geometric representation we introduce the CL-vMF mechanism based on the von Mises-Fisher (vMF) distribution. Meanwhile we introduce the von Mises-Fisher Outlier Detection and Interaction (vMF-ODI) strategy to identify hard examples thus enhancing collaboration between System1 and System2 for complex reasoning realization. Comprehensive evaluation of our proposed ICL demonstrates significant resistance to forgetting and superior performance relative to existing methods. Code is available at github.com/ICL.
****

ZeroNVS: Zero-Shot 360-Degree View Synthesis from a Single Image

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9420-9429

We introduce a 3D-aware diffusion model ZeroNVS for single-image novel view synthesis for in-the-wild scenes. While existing methods are designed for single objects with masked backgrounds we propose new techniques to address challenges introduced by in-the-wild multi-object scenes with complex backgrounds. Specifically we train a generative prior on a mixture of data sources that capture object-centric indoor and outdoor scenes. To address issues from data mixture such as depth-scale ambiguity we propose a novel camera conditioning parameterization and normalization scheme. Further we observe that Score Distillation Sampling (SDS) tends to truncate the distribution of complex backgrounds during distillation of 360-degree scenes and propose "SDS anchoring" to improve the diversity of synthesized novel views. Our model sets a new state-of-the-art result in LPIPS on the DTU dataset in the zero-shot setting even outperforming methods specifically trained on DTU. We further adapt the challenging Mip-NeRF 360 dataset as a new benchmark for single-image novel view synthesis and demonstrate strong performance

in this setting. Code and models will be publicly available.
****
Restoration by Generation with Constrained Priors
Zheng Ding, Xuaner Zhang, Zhuowen Tu, Zhihao Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2567-2577
The inherent generative power of denoising diffusion models makes them well-suited for image restoration tasks where the objective is to find the optimal high-quality image within the generative space that closely resembles the input image. We propose a method to adapt a pretrained diffusion model for image restoration by simply adding noise to the input image to be restored and then denoise. Our method is based on the observation that the space of a generative model needs to be constrained. We impose this constraint by finetuning the generative model with a set of anchor images that capture the characteristics of the input image. With the constrained space we can then leverage the sampling strategy used for generation to do image restoration. We evaluate against previous methods and show superior performances on multiple real-world restoration datasets in preserving identity and image quality. We also demonstrate an important and practical application on personalized restoration where we use a personal album as the anchor images to constrain the generative space. This approach allows us to produce results that accurately preserve high-frequency details which previous works are unable to do. Project webpage: https://gen2res.github.io.
****
Snapshot Lidar: Fourier Embedding of Amplitude and Phase for Single-Image Depth Reconstruction
Sarah Friday, Yunzi Shi, Yaswanth Cherivirala, Vishwanath Saragadam, Adithya Pediredla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25203-25212
Amplitude modulated continuous-wave time-of-flight (AMCW-ToF) cameras are finding applications as flash Lidars in autonomous navigation robotics and AR/VR applications. A conventional CW-ToF camera requires illuminating the scene with a temporally varying light source and demodulating a set of quadrature measurements to recover the scene's depth and intensity. Capturing the four measurements in sequence renders the system slow invariably causing inaccuracies in depth estimates due to motion in the scene or the camera. To mitigate this problem we propose a snapshot Lidar that captures amplitude and phase simultaneously as a single time-of-flight hologram. Uniquely our approach requires minimal changes to existing CW-ToF imaging hardware. To demonstrate the efficacy of the proposed system we design and build a lab prototype and evaluate it under varying scene geometries illumination conditions and compare the reconstructed depth measurements against conventional techniques. We rigorously evaluate the robustness of our system on diverse real-world scenes to show that our technique results in a significant reduction in data bandwidth with minimal loss in reconstruction accuracy. As high-resolution CW-ToF cameras are becoming ubiquitous increasing their temporal resolution by four times enables robust real-time capture of geometries of dynamic scenes.
****
Convolutional Prompting meets Language Models for Continual Learning
Anurag Roy, Riddhiman Moulick, Vinay K. Verma, Saptarshi Ghosh, Abir Das; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23616-23626
Continual Learning (CL) enables machine learning models to learn from continuously shifting new training data in absence of data from old tasks. Recently pre-trained vision transformers combined with prompt tuning have shown promise for overcoming catastrophic forgetting in CL. These approaches rely on a pool of learnable prompts which can be inefficient in sharing knowledge across tasks leading to inferior performance. In addition the lack of fine-grained layer specific prompts does not allow these to fully express the strength of the prompts for CL. We address these limitations by proposing ConvPrompt a novel convolutional prompt creation mechanism that maintains layer-wise shared embeddings enabling both layer-specific learning and better concept transfer across tasks. The intelligent u

se of convolution enables us to maintain a low parameter overhead without compro mising performance. We further leverage Large Language Models to generate fine-g rained text descriptions of each category which are used to get task similarity and dynamically decide the number of prompts to be learned. Extensive experiment s demonstrate the superiority of ConvPrompt and improves SOTA by 3% with signifi cantly less parameter overhead. We also perform strong ablation over various mod ules to disentangle the importance of different components.
****

Blur-aware Spatio-temporal Sparse Transformer for Video Deblurring
Huicong Zhang, Haozhe Xie, Hongxun Yao; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2673-2681
Video deblurring relies on leveraging information from other frames in the video sequence to restore the blurred regions in the current frame. Mainstream approa ches employ bidirectional feature propagation spatio-temporal transformers or a combination of both to extract information from the video sequence. However limi tations in memory and computational resources constraints the temporal window le ngth of the spatio-temporal transformer preventing the extraction of longer temp oral contextual information from the video sequence. Additionally bidirectional feature propagation is highly sensitive to inaccurate optical flow in blurry fra mes leading to error accumulation during the propagation process. To address the se issues we propose BSSTNet Blur-aware Spatio-temporal Sparse Transformer Netwo rk. It introduces the blur map which converts the originally dense attention int o a sparse form enabling a more extensive utilization of information throughout the entire video sequence. Specifically BSSTNet (1) uses a longer temporal windo w in the transformer leveraging information from more distant frames to restore the blurry pixels in the current frame. (2) introduces bidirectional feature pro pagation guided by blur maps which reduces error accumulation caused by the blur frame. The experimental results demonstrate the proposed BSSTNet outperforms th e state-of-the-art methods on the GoPro and DVD datasets.
****

Towards Learning a Generalist Model for Embodied Navigation
Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, Liwei Wang; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1 3624-13634
Building a generalist agent that can interact with the world is an ultimate goal for humans thus spurring the research for embodied navigation where an agent is required to navigate according to instructions or respond to queries. Despite t he major progress attained previous works primarily focus on task-specific agent s and lack generalizability to unseen scenarios. Recently LLMs have presented re markable capabilities across various fields and provided a promising opportunity for embodied navigation. Drawing on this we propose the first generalist model for embodied navigation NaviLLM. It adapts LLMs to embodied navigation by introd ucing schema-based instruction. The schema-based instruction flexibly casts vari ous tasks into generation problems thereby unifying a wide range of tasks. This approach allows us to integrate diverse data sources from various datasets into the training equipping NaviLLM with a wide range of capabilities required by emb odied navigation. We conduct extensive experiments to evaluate the performance a nd generalizability of our model. The experimental results demonstrate that our unified model achieves state-of-the-art performance on CVDN SOON and ScanQA. Spe cifically it surpasses the previous stats-of-the-art method by a significant mar gin of 29% in goal progress on CVDN. Moreover our model also demonstrates strong generalizability and presents impressive results on unseen tasks e.g. embodied question answering and 3D captioning.
****

DiffusionPoser: Real-time Human Motion Reconstruction From Arbitrary Sparse Sens ors Using Autoregressive Diffusion
Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, C. Karen Liu; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2513-2523
Motion capture from a limited number of body-worn sensors such as inertial measu

rement units (IMUs) and pressure insoles has important applications in health hu man performance and entertainment. Recent work has focused on accurately reconst ructing whole-body motion from a specific sensor configuration using six IMUs. W hile a common goal across applications is to use the minimal number of sensors t o achieve required accuracy the optimal arrangement of the sensors might differ from application to application. We propose a single diffusion model DiffusionPo ser which reconstructs human motion in real-time from an arbitrary combination o f sensors including IMUs placed at specified locations and pressure insoles. Unl ike existing methods our model grants users the flexibility to determine the num ber and arrangement of sensors tailored to the specific activity of interest wit hout the need for retraining. A novel autoregressive inferencing scheme ensures real-time motion reconstruction that closely aligns with measured sensor signals . The generative nature of DiffusionPoser ensures realistic behavior even for de grees-of-freedom not directly measured. Qualitative results can be found on our project website.
****
MANUS: Markerless Grasp Capture using Articulated 3D Gaussians
Chandradeep Pokhariya, Ishaan Nikhil Shah, Angela Xing, Zekun Li, Kefan Chen, Av inash Sharma, Srinath Sridhar; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2024, pp. 2197-2208
Understanding how we grasp objects with our hands has important applications in areas like robotics and mixed reality. However this challenging problem requires accurate modeling of the contact between hands and objects.To capture grasps ex isting methods use skeletons meshes or parametric models that does not represent hand shape accurately resulting in inaccurate contacts. We present MANUS a meth od for Markerless Hand-Object Grasp Capture using Articulated 3D Gaussians. We b uild a novel articulated 3D Gaussians representation that extends 3D Gaussian sp latting for high-fidelity representation of articulating hands. Since our repres entation uses Gaussian primitives optimized from the multi-view pixel-aligned lo sses it enables us to efficiently and accurately estimate contacts between the h and and the object. For the most accurate results our method requires tens of ca mera views that current datasets do not provide. We therefore build MANUS-Grasps a new dataset that contains hand-object grasps viewed from 50+ cameras across 3 0+ scenes 3 subjects and comprising over 7M frames. In addition to extensive qua litative results we also show that our method outperforms others on a quantitati ve contact evaluation method that uses paint transfer from the object to the han d.
****
Distilling Semantic Priors from SAM to Efficient Image Restoration Models
Quan Zhang, Xiaoyu Liu, Wei Li, Hanting Chen, Junchao Liu, Jie Hu, Zhiwei Xiong, Chun Yuan, Yunhe Wang; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2024, pp. 25409-25419
In image restoration (IR) leveraging semantic priors from segmentation models ha s been a common approach to improve performance. The recent segment anything mod el (SAM) has emerged as a powerful tool for extracting advanced semantic priors to enhance IR tasks. However the computational cost of SAM is prohibitive for IR compared to existing smaller IR models. The incorporation of SAM for extracting semantic priors considerably hampers the model inference efficiency. To address this issue we propose a general framework to distill SAM's semantic knowledge t o boost exiting IR models without interfering with their inference process. Spec ifically our proposed framework consists of the semantic priors fusion (SPF) sch eme and the semantic priors distillation (SPD) scheme. SPF fuses two kinds of in formation between the restored image predicted by the original IR model and the semantic mask predicted by SAM for the refined restored image. SPD leverages a s elf-distillation manner to distill the fused semantic priors to boost the perfor mance of original IR models. Additionallywe design a semantic-guided relation (S GR) module for SPD which ensures semantic feature representation space consisten cy to fully distill the priors. We demonstrate the effectiveness of our framewor k across multiple IR models and tasks including deraining deblurring and denoisi ng.

****
Learning Intra-view and Cross-view Geometric Knowledge for Stereo Matching
Rui Gong, Weide Liu, Zaiwang Gu, Xulei Yang, Jun Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20752-20762
Geometric knowledge has been shown to be beneficial for the stereo matching task. However prior attempts to integrate geometric insights into stereo matching algorithms have largely focused on geometric knowledge from single images while crucial cross-view factors such as occlusion and matching uniqueness have been overlooked. To address this gap we propose a novel Intra-view and Cross-view Geometric knowledge learning Network (ICGNet) specifically crafted to assimilate both intra-view and cross-view geometric knowledge. ICGNet harnesses the power of interest points to serve as a channel for intra-view geometric understanding. Simultaneously it employs the correspondences among these points to capture cross-view geometric relationships. This dual incorporation empowers the proposed ICGNet to leverage both intra-view and cross-view geometric knowledge in its learning process substantially improving its ability to estimate disparities. Our extensive experiments demonstrate the superiority of the ICGNet over contemporary leading models. The code will be available at https://github.com/DFSDDDDD1199/ICGNet.
****
Rethinking the Evaluation Protocol of Domain Generalization
Han Yu, Xingxuan Zhang, Renzhe Xu, Jiashuo Liu, Yue He, Peng Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21897-21908
Domain generalization aims to solve the challenge of Out-of-Distribution (OOD) generalization by leveraging common knowledge learned from multiple training domains to generalize to unseen test domains. To accurately evaluate the OOD generalization ability it is required that test data information is unavailable. However the current domain generalization protocol may still have potential test data information leakage. This paper examines the risks of test data information leakage from two aspects of the current evaluation protocol: supervised pretraining on ImageNet and oracle model selection. We propose modifications to the current protocol that we should employ self-supervised pretraining or train from scratch instead of employing the current supervised pretraining and we should use multiple test domains. These would result in a more precise evaluation of OOD generalization ability. We also rerun the algorithms with the modified protocol and introduce new leaderboards to encourage future research in domain generalization with a fairer comparison.
****
Aligning Logits Generatively for Principled Black-Box Knowledge Distillation
Jing Ma, Xiang Xiang, Ke Wang, Yuchuan Wu, Yongbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23148-23157
Black-Box Knowledge Distillation (B2KD) is a formulated problem for cloud-to-edge model compression with invisible data and models hosted on the server. B2KD faces challenges such as limited Internet exchange and edge-cloud disparity of data distributions. In this paper we formalize a two-step workflow consisting of deprivatization and distillation and theoretically provide a new optimization direction from logits to cell boundary different from direct logits alignment. With its guidance we propose a new method Mapping-Emulation KD (MEKD) that distills a black-box cumbersome model into a lightweight one. Our method does not differentiate between treating soft or hard responses and consists of: 1) deprivatization: emulating the inverse mapping of the teacher function with a generator and 2) distillation: aligning low-dimensional logits of the teacher and student models by reducing the distance of high-dimensional image points. For different teacher-student pairs our method yields inspiring distillation performance on various benchmarks and outperforms the previous state-of-the-art approaches.
****
BerfScene: Bev-conditioned Equivariant Radiance Fields for Infinite 3D Scene Generation

Qihang Zhang, Yinghao Xu, Yujun Shen, Bo Dai, Bolei Zhou, Ceyuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6839-6849
Generating large-scale 3D scenes cannot simply apply existing 3D object synthesis technique since 3D scenes usually hold complex spatial configurations and consist of a number of objects at varying scales. We thus propose a practical and efficient 3D representation that incorporates an equivariant radiance field with the guidance of a bird's-eye view (BEV) map. Concretely objects of synthesized 3D scenes could be easily manipulated through steering the corresponding BEV maps. Moreover by adequately incorporating positional encoding and low-pass filters into the generator the representation becomes equivariant to the given BEV map. Such equivariance allows us to produce large-scale even infinite-scale 3D scenes via synthesizing local scenes and then stitching them with smooth consistency. Extensive experiments on 3D scene datasets demonstrate the effectiveness of our approach. Our project website is at: https://https://zqh0253.github.io/BerfScene.
****
3D Facial Expressions through Analysis-by-Neural-Synthesis
George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, Petros Maragos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2490-2501
While existing methods for 3D face reconstruction from in-the-wild images excel at recovering the overall face shape they commonly miss subtle extreme asymmetric or rarely observed expressions. We improve upon these methods with SMIRK (Spatial Modeling for Image-based Reconstruction of Kinesics) which faithfully reconstructs expressive 3D faces from images. We identify two key limitations in existing methods: shortcomings in their self-supervised training formulation and a lack of expression diversity in the training images. For training most methods employ differentiable rendering to compare a predicted face mesh with the input image along with a plethora of additional loss functions. This differentiable rendering loss not only has to provide supervision to optimize for 3D face geometry camera albedo and lighting which is an ill-posed optimization problem but the domain gap between rendering and input image further hinders the learning process. Instead SMIRK replaces the differentiable rendering with a neural rendering module that given the rendered predicted mesh geometry and sparsely sampled pixels of the input image generates a face image. As the neural rendering gets color information from sampled image pixels supervising with neural rendering-based reconstruction loss can focus solely on the geometry. Further it enables us to generate images of the input identity with varying expressions while training. These are then utilized as input to the reconstruction model and used as supervision with ground truth geometry. This effectively augments the training data and enhances the generalization for diverse expressions. Our qualitative quantitative and particularly our perceptual evaluations demonstrate that SMIRK achieves the new state-of-the art performance on accurate expression reconstruction. For our method's source code demo video and more please visit our project webpage: https://georgeretsi.github.io/smirk/.
****
HoloVIC: Large-scale Dataset and Benchmark for Multi-Sensor Holographic Intersection and Vehicle-Infrastructure Cooperative
Cong Ma, Lei Qiao, Chengkai Zhu, Kai Liu, Zelong Kong, Qing Li, Xueqi Zhou, Yuheng Kan, Wei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22129-22138
Vehicle-to-everything (V2X) is a popular topic in the field of Autonomous Driving in recent years. Vehicle-infrastructure cooperation (VIC) becomes one of the important research area. Due to the complexity of traffic conditions such as blind spots and occlusion it greatly limits the perception capabilities of single-view roadside sensing systems. To further enhance the accuracy of roadside perception and provide better information to the vehicle side in this paper we constructed holographic intersections with various layouts to build a large-scale multi-sensor holographic vehicle-infrastructure cooperation dataset called HoloVIC. Our dataset includes 3 different types of sensors (Camera Lidar Fisheye) and emplo

ys 4 sensor-layouts based on the different intersections. Each intersection is e quipped with 6-18 sensors to capture synchronous data. While autonomous vehicles pass through these intersections for collecting VIC data. HoloVIC contains in t otal on 100k+ synchronous frames from different sensors. Additionally we annotat ed 3D bounding boxes based on Camera Fisheye and Lidar. We also associate the ID s of the same objects across different devices and consecutive frames in sequenc e. Based on HoloVIC we formulated four tasks to facilitate the development of re lated research. We also provide benchmarks for these tasks.
****
Unleashing the Potential of SAM for Medical Adaptation via Hierarchical Decoding
Zhiheng Cheng, Qingyue Wei, Hongru Zhu, Yan Wang, Liangqiong Qu, Wei Shao, Yuyin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2024, pp. 3511-3522
The Segment Anything Model (SAM) has garnered significant attention for its vers atile segmentation abilities and intuitive prompt-based interface. However its a pplication in medical imaging presents challenges requiring either substantial t raining costs and extensive medical datasets for full model fine-tuning or high-quality prompts for optimal performance. This paper introduces H-SAM: a prompt-f ree adaptation of SAM tailored for efficient fine-tuning of medical images via a two-stage hierarchical decoding procedure. In the initial stage H-SAM employs S AM's original decoder to generate a prior probabilistic mask guiding a more intr icate decoding process in the second stage. Specifically we propose two key desi gns: 1) A class-balanced mask-guided self-attention mechanism addressing the unb alanced label distribution enhancing image embedding; 2) A learnable mask cross-attention mechanism spatially modulating the interplay among different image reg ions based on the prior mask. Moreover the inclusion of a hierarchical pixel dec oder in H-SAM enhances its proficiency in capturing fine-grained and localized d etails. This approach enables SAM to effectively integrate learned medical prior s facilitating enhanced adaptation for medical image segmentation with limited s amples. Our H-SAM demonstrates a 4.78% improvement in average Dice compared to e xisting prompt-free SAM variants for multi-organ segmentation using only 10% of 2D slices. Notably without using any unlabeled data H-SAM even outperforms state -of-the-art semi-supervised models relying on extensive unlabeled training data across various medical datasets. Our code is available at https://github.com/Ccc ccczh404/H-SAM.
****
Puff-Net: Efficient Style Transfer with Pure Content and Style Feature Fusion Ne twork
Sizhe Zheng, Pan Gao, Peng Zhou, Jie Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8059-8068
Style transfer aims to render an image with the artistic features of a style ima ge while maintaining the original structure. Various methods have been put forwa rd for this task but some challenges still exist. For instance it is difficult f or CNN-based methods to handle global information and long-range dependencies be tween input images for which transformer-based methods have been proposed. Altho ugh transformer can better model the relationship between content and style imag es they require high-cost hardware and time-consuming inference. To address thes e issues we design a novel transformer model that includes only encoders thus si gnificantly reducing the computational cost. In addition we also find that exist ing style transfer methods may lead to images under-stylied or missing content. In order to achieve better stylization we design a content feature extractor and a style feature extractor. Then we can feed pure content and style images into the transformer. Finally we propose a network model termed Puff-Net i.e. efficie nt style transfer with pure content and style feature fusion network. Through qu alitative and quantitative experiments we demonstrate the advantages of our mode l compared to state-of-the-art ones in the literature. The code is availabel at https://github.com/ZszYmy9/Puff-Net.
****
Towards Progressive Multi-Frequency Representation for Image Warping
Jun Xiao, Zihang Lyu, Cong Zhang, Yakun Ju, Changjian Shui, Kin-Man Lam; Proceed

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2995-3004

Image warping a classic task in computer vision aims to use geometric transformations to change the appearance of images. Recent methods learn the resampling kernels for warping through neural networks to estimate missing values in irregular grids which however fail to capture local variations in deformed content and produce images with distortion and less high-frequency details. To address this issue this paper proposes an effective method namely MFR to learn Multi-Frequency Representations from input images for image warping. Specifically we propose a progressive filtering network to learn image representations from different frequency subbands and generate deformable images in a coarse-to-fine manner. Furthermore we employ learnable Gabor wavelet filters to improve the model's capability to learn local spatial-frequency representations. Comprehensive experiments including homography transformation equirectangular to perspective projection and asymmetric image super-resolution demonstrate that the proposed MFR significantly outperforms state-of-the-art image warping methods. Our method also showcases superior generalization to out-of-distribution domains where the generated images are equipped with rich details and less distortion thereby high visual quality. The source code is available at https://github.com/junxiao01/MFR.

****

Learning to Control Camera Exposure via Reinforcement Learning

Kyunghyun Lee, Ukcheol Shin, Byeong-Uk Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2975-2983

Adjusting camera exposure in arbitrary lighting conditions is the first step to ensure the functionality of computer vision applications. Poorly adjusted camera exposure often leads to critical failure and performance degradation. Traditional camera exposure control methods require multiple convergence steps and time-consuming processes making them unsuitable for dynamic lighting conditions. In this paper we propose a new camera exposure control framework that rapidly controls camera exposure while performing real-time processing by exploiting deep reinforcement learning. The proposed framework consists of four contributions: 1) a simplified training ground to simulate real-world's diverse and dynamic lighting changes 2) flickering and image attribute-aware reward design along with lightweight state design for real-time processing 3) a static-to-dynamic lighting curriculum to gradually improve the agent's exposure-adjusting capability and 4) domain randomization techniques to alleviate the limitation of the training ground and achieve seamless generalization in the wild. As a result our proposed method rapidly reaches a desired exposure level within five steps with real-time processing (1 ms). Also the acquired images are well-exposed and show superiority in various computer vision tasks such as feature extraction and object detection.

****

Splatter Image: Ultra-Fast Single-View 3D Reconstruction

Stanislaw Szymanowicz, Chrisitian Rupprecht, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10208-10217

We introduce the Splatter Image an ultra-efficient approach for monocular 3D object reconstruction. Splatter Image is based on Gaussian Splatting which allows fast and high-quality reconstruction of 3D scenes from multiple images. We apply Gaussian Splatting to monocular reconstruction by learning a neural network that at test time performs reconstruction in a feed-forward manner at 38 FPS. Our main innovation is the surprisingly straightforward design of this network which using 2D operators maps the input image to one 3D Gaussian per pixel. The resulting set of Gaussians thus has the form an image the Splatter Image. We further extend the method take several images as input via cross-view attention. Owning to the speed of the renderer (588 FPS) we use a single GPU for training while generating entire images at each iteration to optimize perceptual metrics like LPIPS. On several synthetic real multi-category and large-scale benchmark datasets we achieve better results in terms of PSNR LPIPS and other metrics while training and evaluating much faster than prior works. Code models and more results are available at https://szymanowiczs.github.io/ splatter-image.

****
Modeling Collaborator: Enabling Subjective Vision Classification With Minimal Hu man Effort via LLM Tool-Use
Imad Eddine Toubal, Aditya Avinash, Neil Gordon Alldrin, Jan Dlabal, Wenlei Zhou , Enming Luo, Otilia Stretcu, Hao Xiong, Chun-Ta Lu, Howard Zhou, Ranjay Krishna , Ariel Fuxman, Tom Duerig; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2024, pp. 17553-17563
From content moderation to wildlife conservation the number of applications that require models to recognize nuanced or subjective visual concepts is growing. T raditionally developing classifiers for such concepts requires substantial manua l effort measured in hours days or even months to identify and annotate data nee ded for training. Even with recently proposed Agile Modeling techniques which en able rapid bootstrapping of image classifiers users are still required to spend 30 minutes or more of monotonous repetitive data labeling just to train a single classifier. Drawing on Fiske's Cognitive Miser theory we propose a new framewor k that alleviates manual effort by replacing human labeling with natural languag e interactions reducing the total effort required to define a concept by an orde r of magnitude: from labeling 2000 images to only 100 plus some natural language interactions. Our framework leverages recent advances in foundation models both large language models and vision-language models to carve out the concept space through conversation and by automatically labeling training data points. Most i mportantly our framework eliminates the need for crowd-sourced annotations. More over our framework ultimately produces lightweight classification models that ar e deployable in cost-sensitive scenarios. Across 15 subjective concepts and acro ss 2 public image classification datasets our trained models outperform traditio nal Agile Modeling as well as state-of-the-art zero-shot classification models l ike ALIGN CLIP CuPL and large visual question answering models like PaLI-X.
****
RNb-NeuS: Reflectance and Normal-based Multi-View 3D Reconstruction
Baptiste Brument, Robin Bruneau, Yvain Quéau, Jean Mélou, François Bernard Lauze , Jean-Denis Durou, Lilian Calvet; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2024, pp. 5230-5239
This paper introduces a versatile paradigm for integrating multi-view reflectanc e (optional) and normal maps acquired through photometric stereo. Our approach e mploys a pixel-wise joint re-parameterization of reflectance and normal consider ing them as a vector of radiances rendered under simulated varying illumination. This re-parameterization enables the seamless integration of reflectance and no rmal maps as input data in neural volume rendering-based 3D reconstruction while preserving a single optimization objective. In contrast recent multi-view photo metric stereo (MVPS) methods depend on multiple potentially conflicting objectiv es. Despite its apparent simplicity our proposed approach outperforms state-of-t he-art approaches in MVPS benchmarks across F-score Chamfer distance and mean an gular error metrics. Notably it significantly improves the detailed 3D reconstru ction of areas with high curvature or low visibility.
****
LOTUS: Evasive and Resilient Backdoor Attacks through Sub-Partitioning
Siyuan Cheng, Guanhong Tao, Yingqi Liu, Guangyu Shen, Shengwei An, Shiwei Feng, Xiangzhe Xu, Kaiyuan Zhang, Shiqing Ma, Xiangyu Zhang; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24798 -24809
Backdoor attack poses a significant security threat to Deep Learning application s. Existing attacks are often not evasive to established backdoor detection tech niques. This susceptibility primarily stems from the fact that these attacks typ ically leverage a universal trigger pattern or transformation function such that the trigger can cause misclassification for any input. In response to this rece nt papers have introduced attacks using sample-specific invisible triggers craft ed through special transformation functions. While these approaches manage to ev ade detection to some extent they reveal vulnerability to existing backdoor miti gation techniques. To address and enhance both evasiveness and resilience we int roduce a novel backdoor attack LOTUS. Specifically it leverages a secret functio

n to separate samples in the victim class into a set of partitions and applies u
nique triggers to different partitions. Furthermore LOTUS incorporates an effect
ive trigger focusing mechanism ensuring only the trigger corresponding to the pa
rtition can induce the backdoor behavior. Extensive experimental results show th
at LOTUS can achieve high attack success rate across 4 datasets and 7 model stru
ctures and effectively evading 13 backdoor detection and mitigation techniques.
The code is available at https://github.com/Megum1/LOTUS.
****
GeoReF: Geometric Alignment Across Shape Variation for Category-level Object Pos
e Refinement
Linfang Zheng, Tze Ho Elden Tse, Chen Wang, Yinghan Sun, Hua Chen, Ales Leonardi
s, Wei Zhang, Hyung Jin Chang; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2024, pp. 10693-10703
Object pose refinement is essential for robust object pose estimation. Previous
work has made significant progress towards instance-level object pose refinement
. Yet category-level pose refinement is a more challenging problem due to large
shape variations within a category and the discrepancies between the target obje
ct and the shape prior. To address these challenges we introduce a novel archite
cture for category-level object pose refinement. Our approach integrates an HS-l
ayer and learnable affine transformations which aims to enhance the extraction a
nd alignment of geometric information. Additionally we introduce a cross-cloud t
ransformation mechanism that efficiently merges diverse data sources. Finally we
 push the limits of our model by incorporating the shape prior information for t
ranslation and size error prediction. We conducted extensive experiments to demo
nstrate the effectiveness of the proposed framework. Through extensive quantitat
ive experiments we demonstrate significant improvement over the baseline method
by a large margin across all metrics.
****
LAN: Learning to Adapt Noise for Image Denoising
Changjin Kim, Tae Hyun Kim, Sungyong Baik; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25193-25202
Removing noise from images a.k.a image denoising can be a very challenging task
since the type and amount of noise can greatly vary for each image due to many f
actors including a camera model and capturing environments. While there have bee
n striking improvements in image denoising with the emergence of advanced deep l
earning architectures and real-world datasets recent denoising networks struggle
 to maintain performance on images with noise that has not been seen during trai
ning. One typical approach to address the challenge would be to adapt a denoisin
g network to new noise distribution. Instead in this work we shift our attention
 to the input noise itself for adaptation rather than adapting a network. Thus w
e keep a pretrained network frozen and adapt an input noise to capture the fine-
grained deviations. As such we propose a new denoising algorithm dubbed Learning
-to-Adapt-Noise (LAN) where a learnable noise offset is directly added to a give
n noisy image to bring a given input noise closer towards the noise distribution
 a denoising network is trained to handle. Consequently the proposed framework e
xhibits performance improvement on images with unseen noise displaying the poten
tial of the proposed research direction.
****
Scaling Up Dynamic Human-Scene Interaction Modeling
Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu
Liu, Yixin Zhu, Siyuan Huang; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2024, pp. 1737-1747
Confronting the challenges of data scarcity and advanced motion synthesis in hum
an-scene interaction modeling we introduce the TRUMANS dataset alongside a novel
 HSI motion synthesis method. TRUMANS stands as the most comprehensive motion-ca
ptured HSI dataset currently available encompassing over 15 hours of human inter
actions across 100 indoor scenes. It intricately captures whole-body human motio
ns and part-level object dynamics focusing on the realism of contact. This datas
et is further scaled up by transforming physical environments into exact virtual
 models and applying extensive augmentations to appearance and motion for both h

umans and objects while maintaining interaction fidelity. Utilizing TRUMANS we d
evise a diffusion-based autoregressive model that efficiently generates HSI sequ
ences of any length taking into account both scene context and intended actions.
 In experiments our approach shows remarkable zero-shot generalizability on a ra
nge of 3D scene datasets (e.g. PROX Replica ScanNet ScanNet++) producing motions
 that closely mimic original motion-captured sequences as confirmed by quantitat
ive experiments and human studies.
****
Semantic-aware SAM for Point-Prompted Instance Segmentation
Zhaoyang Wei, Pengfei Chen, Xuehui Yu, Guorong Li, Jianbin Jiao, Zhenjun Han; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 3585-3594
Single-point annotation in visual tasks with the goal of minimizing labeling cos
ts is becoming increasingly prominent in research. Recently visual foundation mo
dels such as Segment Anything (SAM) have gained widespread usage due to their ro
bust zero-shot capabilities and exceptional annotation performance. However SAM'
s class-agnostic output and high confidence in local segmentation introduce sema
ntic ambiguity posing a challenge for precise category-specific segmentation. In
 this paper we introduce a cost-effective category-specific segmenter using SAM.
 To tackle this challenge we have devised a Semantic-Aware Instance Segmentation
 Network (SAPNet) that integrates Multiple Instance Learning (MIL) with matching
 capability and SAM with point prompts. SAPNet strategically selects the most re
presentative mask proposals generated by SAM to supervise segmentation with a sp
ecific focus on object category information. Moreover we introduce the Point Dis
tance Guidance and Box Mining Strategy to mitigate inherent challenges: group an
d local issues in weakly supervised segmentation. These strategies serve to furt
her enhance the overall segmentation performance. The experimental results on Pa
scal VOC and COCO demonstrate the promising performance of our proposed SAPNet e
mphasizing its semantic matching capabilities and its potential to advance point
-prompted instance segmentation. The code is available at https://github.com/zha
oyangwei123/SAPNet.
****
Learning Group Activity Features Through Person Attribute Prediction
Chihiro Nakatani, Hiroaki Kawashima, Norimichi Ukita; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18233-
18242
This paper proposes Group Activity Feature (GAF) learning in which features of m
ulti-person activity are learned as a compact latent vector. Unlike prior work i
n which the manual annotation of group activities is required for supervised lea
rning our method learns the GAF through person attribute prediction without grou
p activity annotations. By learning the whole network in an end-to-end manner so
 that the GAF is required for predicting the person attributes of people in a gr
oup the GAF is trained as the features of multi-person activity. As a person att
ribute we propose to use a person's action class and appearance features because
 the former is easy to annotate due to its simpleness and the latter requires no
 manual annotation. In addition we introduce a location-guided attribute predict
ion to disentangle the complex GAF for extracting the features of each target pe
rson properly. Various experimental results validate that our method outperforms
 SOTA methods quantitatively and qualitatively on two public datasets. Visualiza
tion of our GAF also demonstrates that our method learns the GAF representing fi
ned-grained group activity classes. Code: https://github.com/chihina/GAFL-CVPR20
24.
****
HUNTER: Unsupervised Human-centric 3D Detection via Transferring Knowledge from
Synthetic Instances to Real Scenes
Yichen Yao, Zimo Jiang, Yujing Sun, Zhencai Zhu, Xinge Zhu, Runnan Chen, Yuexin
Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2024, pp. 28120-28129
Human-centric 3D scene understanding has recently drawn increasing attention dri
ven by its critical impact on robotics. However human-centric real-life scenario

s are extremely diverse and complicated and humans have intricate motions and interactions. With limited labeled data supervised methods are difficult to generalize to general scenarios hindering real-life applications. Mimicking human intelligence we propose an unsupervised 3D detection method for human-centric scenarios by transferring the knowledge from synthetic human instances to real scenes. To bridge the gap between the distinct data representations and feature distributions of synthetic models and real point clouds we introduce novel modules for effective instance-to-scene representation transfer and synthetic-to-real feature alignment. Remarkably our method exhibits superior performance compared to current state-of-the-art techniques achieving 87.8% improvement in mAP and closely approaching the performance of fully supervised methods (62.15 mAP vs. 69.02 mAP) on HuCenLife Dataset.
****
Improving Transferable Targeted Adversarial Attacks with Model Self-Enhancement
Han Wu, Guanyan Ou, Weibin Wu, Zibin Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24615-24624
Various transfer attack methods have been proposed to evaluate the robustness of deep neural networks (DNNs). Although manifesting remarkable performance in generating untargeted adversarial perturbations existing proposals still fail to achieve high targeted transferability. In this work we discover that the adversarial perturbations' overfitting towards source models of mediocre generalization capability can hurt their targeted transferability. To address this issue we focus on enhancing the source model's generalization capability to improve its ability to conduct transferable targeted adversarial attacks. In pursuit of this goal we propose a novel model self-enhancement method that incorporates two major components: Sharpness-Aware Self-Distillation (SASD) and Weight Scaling (WS). Specifically SASD distills a fine-tuned auxiliary model which mirrors the source model's structure into the source model while flattening the source model's loss landscape. WS obtains an approximate ensemble of numerous pruned models to perform model augmentation which can be conveniently synergized with SASD to elevate the source model's generalization capability and thus improve the resultant targeted perturbations' transferability. Extensive experiments corroborate the effectiveness of the proposed method. Notably under the black-box setting our approach can outperform the state-of-the-art baselines by a significant margin of 12.2% on average in terms of the obtained targeted transferability. Code is available at https://github.com/g4alllf/SASD.
****
Unsupervised Learning of Category-Level 3D Pose from Object-Centric Videos
Leonhard Sommer, Artur Jesslen, Eddy Ilg, Adam Kortylewski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22787-22796
Category-level 3D pose estimation is a fundamentally important problem in computer vision and robotics e.g. for embodied agents or to train 3D generative models. However so far methods that estimate the category-level object pose require either large amounts of human annotations CAD models or input from RGB-D sensors. In contrast we tackle the problem of learning to estimate the category-level 3D pose only from casually taken object-centric videos without human supervision. We propose a two-step pipeline: First we introduce a multi-view alignment procedure that determines canonical camera poses across videos with a novel and robust cyclic distance formulation for geometric and appearance matching using reconstructed coarse meshes and DINOv2 features. In a second step the canonical poses and reconstructed meshes enable us to train a model for 3D pose estimation from a single image. In particular our model learns to estimate dense correspondences between images and a prototypical 3D template by predicting for each pixel in a 2D image a feature vector of the corresponding vertex in the template mesh. We demonstrate that our method outperforms all baselines at the unsupervised alignment of object-centric videos by a large margin and provides faithful and robust predictions in-the-wild on the Pascal3D+ and ObjectNet3D datasets.
****
Plug-and-Play Diffusion Distillation

Yi-Ting Hsiao, Siavash Khodadadeh, Kevin Duarte, Wei-An Lin, Hui Qu, Mingi Kwon, Ratheesh Kalarot; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13743-13752

Diffusion models have shown tremendous results in image generation. However due to the iterative nature of the diffusion process and its reliance on classifier-free guidance inference times are slow. In this paper we propose a new distillation approach for guided diffusion models in which an external lightweight guide model is trained while the original text-to-image model remains frozen. We show that our method reduces the inference computation of classifier-free guided latent-space diffusion models by almost half and only requires 1% trainable parameters of the base model. Furthermore once trained our guide model can be applied to various fine-tuned domain-specific versions of the base diffusion model without the need for additional training: this "plug-and-play" functionality drastically improves inference computation while maintaining the visual fidelity of generated images. Empirically we show that our approach is able to produce visually appealing results and achieve a comparable FID score to the teacher with as few as 8 to 16 steps.

****

MindBridge: A Cross-Subject Brain Decoding Framework

Shizun Wang, Songhua Liu, Zhenxiong Tan, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11333-11342

Brain decoding a pivotal field in neuroscience aims to reconstruct stimuli from acquired brain signals primarily utilizing functional magnetic resonance imaging (fMRI). Currently brain decoding is confined to a per-subject-per-model paradigm limiting its applicability to the same individual for whom the decoding model is trained. This constraint stems from three key challenges: 1) the inherent variability in input dimensions across subjects due to differences in brain size; 2) the unique intrinsic neural patterns influencing how different individuals perceive and process sensory information; 3) limited data availability for new subjects in real-world scenarios hampers the performance of decoding models. In this paper we present a novel approach MindBridge that achieves cross-subject brain decoding by employing only one model. Our proposed framework establishes a generic paradigm capable of addressing these challenges by introducing biological-inspired aggregation function and novel cyclic fMRI reconstruction mechanism for subject-invariant representation learning. Notably by cycle reconstruction of fMRI MindBridge can enable novel fMRI synthesis which also can serve as pseudo data augmentation. Within the framework we also devise a novel reset-tuning method for adapting a pretrained model to a new subject. Experimental results demonstrate MindBridge's ability to reconstruct images for multiple subjects which is competitive with dedicated subject-specific models. Furthermore with limited data for a new subject we achieve a high level of decoding accuracy surpassing that of subject-specific models. This advancement in cross-subject brain decoding suggests promising directions for wider applications in neuroscience and indicates potential for more efficient utilization of limited fMRI data in real-world scenarios. Project page: https://littlepure2333.github.io/MindBridge

****

Make Pixels Dance: High-Dynamic Video Generation

Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, Hang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8850-8860

Creating high-dynamic videos such as motion-rich actions and sophisticated visual effects poses a significant challenge in the field of artificial intelligence. Unfortunately current state-of-the-art video generation methods primarily focusing on text-to-video generation tend to produce video clips with minimal motions despite maintaining high fidelity. We argue that relying solely on text instructions is insufficient and suboptimal for video generation. In this paper we introduce PixelDance a novel approach based on diffusion models that incorporates image instructions for both the first and last frames in conjunction with text instructions for video generation. Comprehensive experimental results demonstrate t

hat PixelDance trained with public data exhibits significantly better proficienc y in synthesizing videos with complex scenes and intricate motions setting a new standard for video generation.
****
MM-Narrator: Narrating Long-form Videos with Multimodal In-Context Learning
Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, Lijuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13647-13657
We present MM-Narrator a novel system leveraging GPT-4 with multimodal in-context learning for the generation of audio descriptions (AD). Unlike previous methods that primarily focused on downstream fine-tuning with short video clips MM-Narrator excels in generating precise audio descriptions for videos of extensive lengths even beyond hours in an autoregressive manner. This capability is made possible by the proposed memory-augmented generation process which effectively utilizes both the short-term textual context and long-term visual memory through an efficient register-and-recall mechanism. These contextual memories compile pertinent past information including storylines and character identities ensuring an accurate tracking and depicting of story-coherent and character-centric audio descriptions. Maintaining the training-free design of MM-Narrator we further propose a complexity-based demonstration selection strategy to largely enhance its multi-step reasoning capability via few-shot multimodal in-context learning (MM-ICL). Experimental results on MAD-eval dataset demonstrate that MM-Narrator consistently outperforms both the existing fine-tuning-based approaches and LLM-based approaches in most scenarios as measured by standard evaluation metrics. Additionally we introduce the first segment-based evaluator for recurrent text generation. Empowered by GPT-4 this evaluator comprehensively reasons and marks AD generation performance in various extendable dimensions.
****
Morphable Diffusion: 3D-Consistent Diffusion for Single-image Avatar Creation
Xiyi Chen, Marko Mihajlovic, Shaofei Wang, Sergey Prokudin, Siyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10359-10370
Recent advances in generative diffusion models have enabled the previously unfeasible capability of generating 3D assets from a single input image or a text prompt. In this work we aim to enhance the quality and functionality of these models for the task of creating controllable photorealistic human avatars. We achieve this by integrating a 3D morphable model into the state-of-the-art multi-view-consistent diffusion approach. We demonstrate that accurate conditioning of a generative pipeline on the articulated 3D model enhances the baseline model performance on the task of novel view synthesis from a single image. More importantly this integration facilitates a seamless and accurate incorporation of facial expression and body pose control into the generation process. To the best of our knowledge our proposed framework is the first diffusion model to enable the creation of fully 3D-consistent animatable and photorealistic human avatars from a single image of an unseen subject; extensive quantitative and qualitative evaluations demonstrate the advantages of our approach over existing state-of-the-art avatar creation models on both novel view and novel expression synthesis tasks. The code for our project is publicly available.
****
Fully Convolutional Slice-to-Volume Reconstruction for Single-Stack MRI
Sean I. Young, Yael Balbastre, Bruce Fischl, Polina Golland, Juan Eugenio Iglesias; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11535-11545
In magnetic resonance imaging (MRI) slice-to-volume reconstruction (SVR) refers to computational reconstruction of an unknown 3D magnetic resonance volume from stacks of 2D slices corrupted by motion. While promising current SVR methods require multiple slice stacks for accurate 3D reconstruction leading to long scans and limiting their use in time-sensitive applications such as fetal fMRI. Here we propose a SVR method that overcomes the shortcomings of previous work and produces state-of-the-art reconstructions in the presence of extreme inter-slice mot

ion. Inspired by the recent success of single-view depth estimation methods we f
ormulate SVR as a single-stack motion estimation task and train a fully convolut
ional network to predict a motion stack for a given slice stack producing a 3D r
econstruction as a byproduct of the predicted motion. Extensive experiments on t
he SVR of adult and fetal brains demonstrate that our fully convolutional method
 is twice as accurate as previous SVR methods. Our code is available at github.c
om/seannz/svr.
****
Enhance Image Classification via Inter-Class Image Mixup with Diffusion Model
Zhicai Wang, Longhui Wei, Tan Wang, Heyu Chen, Yanbin Hao, Xiang Wang, Xiangnan
He, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2024, pp. 17223-17233
Text-to-image (T2I) generative models have recently emerged as a powerful tool e
nabling the creation of photo-realistic images and giving rise to a multitude of
 applications. However the effective integration of T2I models into fundamental
image classification tasks remains an open question. A prevalent strategy to bol
ster image classification performance is through augmenting the training set wit
h synthetic images generated by T2I models. In this study we scrutinize the shor
tcomings of both current generative and conventional data augmentation technique
s. Our analysis reveals that these methods struggle to produce images that are b
oth faithful (in terms of foreground objects) and diverse (in terms of backgroun
d contexts) for domain-specific concepts. To tackle this challenge we introduce
an innovative inter-class data augmentation method known as Diff-Mix (\href http
s://github.com/Zhicaiwww/Diff-Mix) https://github.com/Zhicaiwww/Diff-Mix   which
 enriches the dataset by performing image translations between classes. Our empi
rical results demonstrate that Diff-Mix achieves a better balance between faithf
ulness and diversity leading to a marked improvement in performance across diver
se image classification scenarios including few-shot conventional and long-tail
classifications for domain-specific datasets.
****
A&B BNN: Add&Bit-Operation-Only Hardware-Friendly Binary Neural Network
Ruichen Ma, Guanchao Qiao, Yian Liu, Liwei Meng, Ning Ning, Yang Liu, Shaogang H
u; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2024, pp. 5704-5713
Binary neural networks utilize 1-bit quantized weights and activations to reduce
 both the model's storage demands and computational burden. However advanced bin
ary architectures still incorporate millions of inefficient and nonhardware-frie
ndly full-precision multiplication operations. A&B BNN is proposed to directly r
emove part of the multiplication operations in a traditional BNN and replace the
 rest with an equal number of bit operations introducing the mask layer and the
quantized RPReLU structure based on the normalizer-free network architecture. Th
e mask layer can be removed during inference by leveraging the intrinsic charact
eristics of BNN with straightforward mathematical transformations to avoid the a
ssociated multiplication operations. The quantized RPReLU structure enables more
 efficient bit operations by constraining its slope to be integer powers of 2. E
xperimental results achieved 92.30% 69.35% and 66.89% on the CIFAR-10 CIFAR-100
and ImageNet datasets respectively which are competitive with the state-of-the-a
rt. Ablation studies have verified the efficacy of the quantized RPReLU structur
e leading to a 1.14% enhancement on the ImageNet compared to using a fixed slope
 RLeakyReLU. The proposed add&bit-operation-only BNN offers an innovative approa
ch for hardware-friendly network architecture.
****
Alpha-CLIP: A CLIP Model Focusing on Wherever You Want
Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dah
ua Lin, Jiaqi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2024, pp. 13019-13029
Contrastive Language-Image Pre-training (CLIP) plays an essential role in extrac
ting valuable content information from images across diverse tasks. It aligns te
xtual and visual modalities to comprehend the entire image including all the det
ails even those irrelevant to specific tasks. However for a finer understanding

and controlled editing of images it becomes crucial to focus on specific regions of interest which can be indicated as points masks or boxes by humans or perception models. To fulfill the requirements we introduce Alpha-CLIP an enhanced version of CLIP with an auxiliary alpha channel to suggest attentive regions and fine-tuned with constructed millions of RGBA region-text pairs. Alpha-CLIP not only preserves the visual recognition ability of CLIP but also enables precise control over the emphasis of image contents. It demonstrates effectiveness in various tasks including but not limited to open-world recognition multimodal large language models and conditional 2D / 3D generation. It has a strong potential to serve as a versatile tool for image-related tasks.
****
FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations
Christian Diller, Thomas Funkhouser, Angela Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19902-19914
We present a generative approach to forecast long-term future human behavior in 3D requiring only weak supervision from readily available 2D human action data. This is a fundamental task enabling many downstream applications. The required ground-truth data is hard to capture in 3D (mocap suits expensive setups) but easy to acquire in 2D (simple RGB cameras). Thus we design our method to only require 2D RGB data at inference time while being able to generate 3D human motion sequences. We use a differentiable 2D projection scheme in an autoregressive manner for weak supervision and an adversarial loss for 3D regularization. Our method predicts long and complex human behavior sequences (e.g. cooking assembly) consisting of multiple sub-actions. We tackle this in a semantically hierarchical manner jointly predicting high-level coarse action labels together with their low-level fine-grained realizations as characteristic 3D human poses. We observe that these two action representations are coupled in nature and joint prediction benefits both action and pose forecasting. Our experiments demonstrate the complementary nature of joint action and 3D pose prediction: our joint approach outperforms each task treated individually enables robust longer-term sequence prediction and improves over alternative approaches to forecast actions and characteristic 3D poses.
****
NightCC: Nighttime Color Constancy via Adaptive Channel Masking
Shuwei Li, Robby T. Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25522-25531
Nighttime conditions pose a significant challenge to color constancy due to the diversity of lighting conditions and the presence of substantial low-light noise. Existing color constancy methods struggle with nighttime scenes frequently leading to imprecise light color estimations. To tackle nighttime color constancy we propose a novel unsupervised domain adaptation approach that utilizes labeled daytime data to facilitate learning on unlabeled nighttime images. To specifically address the unique lighting conditions of nighttime and ensure the robustness of pseudo labels we propose adaptive channel masking and light uncertainty. By selectively masking channels that are less sensitive to lighting conditions adaptive channel masking directs the model to progressively focus on features less affected by variations in light colors and noise. Additionally our model leverages light uncertainty to provide a pixel-wise uncertainty estimation regarding light color prediction which helps avoid learning from incorrect labels. Our model demonstrates a significant improvement in accuracy achieving 21.5% lower Mean Angular Error (MAE) compared to the state-of-the-art method on our nighttime dataset.
****
Task-aligned Part-aware Panoptic Segmentation through Joint Object-Part Representations
Daan de Geus, Gijs Dubbelman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3174-3183
Part-aware panoptic segmentation (PPS) requires (a) that each foreground object and background region in an image is segmented and classified and (b) that all p

arts within foreground objects are segmented classified and linked to their pare nt object. Existing methods approach PPS by separately conducting object-level a nd part-level segmentation. However their part-level predictions are not linked to individual parent objects. Therefore their learning objective is not aligned with the PPS task objective which harms the PPS performance. To solve this and m ake more accurate PPS predictions we propose Task-Aligned Part-aware Panoptic Se gmentation (TAPPS). This method uses a set of shared queries to jointly predict (a) object-level segments and (b) the part-level segments within those same obje cts. As a result TAPPS learns to predict part-level segments that are linked to individual parent objects aligning the learning objective with the task objectiv e and allowing TAPPS to leverage joint object-part representations. With experim ents we show that TAPPS considerably outperforms methods that predict objects an d parts separately and achieves new state-of-the-art PPS results.
****
From Activation to Initialization: Scaling Insights for Optimizing Neural Fields
Hemanth Saratchandran, Sameera Ramasinghe, Simon Lucey; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 413-422
In the realm of computer vision Neural Fields have gained prominence as a contem porary tool harnessing neural networks for signal representation. Despite the re markable progress in adapting these networks to solve a variety of problems the field still lacks a comprehensive theoretical framework. This article aims to ad dress this gap by delving into the intricate interplay between initialization an d activation providing a foundational basis for the robust optimization of Neura l Fields. Our theoretical insights reveal a deep-seated connection among network initialization architectural choices and the optimization process emphasizing t he need for a holistic approach when designing cutting-edge Neural Fields.
****
UnScene3D: Unsupervised 3D Instance Segmentation for Indoor Scenes
David Rozenberszki, Or Litany, Angela Dai; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19957-19967
3D instance segmentation is fundamental to geometric understanding of the world around us. Existing methods for instance segmentation of 3D scenes rely on super vision from expensive manual 3D annotations. We propose UnScene3D the first full y unsupervised 3D learning approach for class-agnostic 3D instance segmentation of indoor scans. UnScene3D first generates pseudo masks by leveraging self-super vised color and geometry features to find potential object regions. We operate o n a basis of 3D segment primitives enabling efficient representation and learnin g on high-resolution 3D data. The coarse proposals are then refined through self -training our model on its predictions. Our approach improves over state-of-the-art unsupervised 3D instance segmentation methods by more than 300% Average Prec ision score demonstrating effective instance segmentation even in challenging cl uttered 3D scenes.
****
Nearest is Not Dearest: Towards Practical Defense against Quantization-condition ed Backdoor Attacks
Boheng Li, Yishuo Cai, Haowei Li, Feng Xue, Zhifeng Li, Yiming Li; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 24523-24533
Model quantization is widely used to compress and accelerate deep neural network s. However recent studies have revealed the feasibility of weaponizing model qua ntization via implanting quantization-conditioned backdoors (QCBs). These specia l backdoors stay dormant on released full-precision models but will come into ef fect after standard quantization. Due to the peculiarity of QCBs existing defens es have minor effects on reducing their threats or are even infeasible. In this paper we conduct the first in-depth analysis of QCBs. We reveal that the activat ion of existing QCBs primarily stems from the nearest rounding operation and is closely related to the norms of neuron-wise truncation errors (i.e. the differen ce between the continuous fullprecision weights and its quantized version). Moti vated by these insights we propose Error-guided Flipped Rounding with Activation

Preservation (EFRAP) an effective and practical defense against QCBs. Specifically EFRAP learns a non-nearest rounding strategy with neuron-wise error norm and layer-wise activation preservation guidance flipping the rounding strategies of neurons crucial for backdoor effects but with minimal impact on clean accuracy. Extensive evaluations on benchmark datasets demonstrate that our EFRAP can defeat state-of-the-art QCB attacks under various settings. Code is available here.
****
DiffAvatar: Simulation-Ready Garment Optimization with Differentiable Simulation
Yifei Li, Hsiao-yu Chen, Egor Larionov, Nikolaos Sarafianos, Wojciech Matusik, Tuur Stuyck; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4368-4378
The realism of digital avatars is crucial in enabling telepresence applications with self-expression and customization. While physical simulations can produce realistic motions for clothed humans they require high-quality garment assets with associated physical parameters for cloth simulations. However manually creating these assets and calibrating their parameters is labor-intensive and requires specialized expertise. Current methods focus on reconstructing geometry but don't generate complete assets for physics-based applications. To address this gap we propose DiffAvatar a novel approach that performs body and garment co-optimization using differentiable simulation. By integrating physical simulation into the optimization loop and accounting for the complex nonlinear behavior of cloth and its intricate interaction with the body our framework recovers body and garment geometry and extracts important material parameters in a physically plausible way. Our experiments demonstrate that our approach generates realistic clothing and body shape suitable for downstream applications. We provide additional insights and results on our webpage: people.csail.mit.edu/liyifei/publication/diffavatar.
****
AlignSAM: Aligning Segment Anything Model to Open Context via Reinforcement Learning
Duojun Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, Guanbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3205-3215
Powered by massive curated training data Segment Anything Model (SAM) has demonstrated its impressive generalization capabilities in open-world scenarios with the guidance of prompts. However the vanilla SAM is class-agnostic and heavily relies on user-provided prompts to segment objects of interest. Adapting this method to diverse tasks is crucial for accurate target identification and to avoid suboptimal segmentation results. In this paper we propose a novel framework termed AlignSAM designed for automatic prompting for aligning SAM to an open context through reinforcement learning. Anchored by an agent AlignSAM enables the generality of the SAM model across diverse downstream tasks while keeping its parameters frozen. Specifically AlignSAM initiates a prompting agent to iteratively refine segmentation predictions by interacting with the foundational model. It integrates a reinforcement learning policy network to provide informative prompts to the foundational models. Additionally a semantic recalibration module is introduced to provide fine-grained labels of prompts enhancing the model's proficiency in handling tasks encompassing explicit and implicit semantics. Experiments conducted on various challenging segmentation tasks among existing foundation models demonstrate the superiority of the proposed AlignSAM over state-of-the-art approaches. Project page: https://github.com/Duojun-Huang/AlignSAM-CVPR2024.
****
A Simple Recipe for Language-guided Domain Generalized Segmentation
Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, Raoul de Charette; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23428-23437
Generalization to new domains not seen during training is one of the long-standing challenges in deploying neural networks in real-world applications. Existing generalization techniques either necessitate external images for augmentation and/or aim at learning invariant representations by imposing various alignment con

straints. Large-scale pretraining has recently shown promising generalization ca
pabilities along with the potential of binding different modalities. For instanc
e the advent of vision-language models like CLIP has opened the doorway for visi
on models to exploit the textual modality. In this paper we introduce a simple f
ramework for generalizing semantic segmentation networks by employing language a
s the source of randomization. Our recipe comprises three key ingredients: (i) t
he preservation of the intrinsic CLIP robustness through minimal fine-tuning (ii
) language-driven local style augmentation and (iii) randomization by locally mi
xing the source and augmented styles during training. Extensive experiments repo
rt state-of-the-art results on various generalization benchmarks.
****
Learning Spatial Adaptation and Temporal Coherence in Diffusion Models for Video
 Super-Resolution
Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, Tao Me
i; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2024, pp. 9232-9241
Diffusion models are just at a tipping point for image super-resolution task. Ne
vertheless it is not trivial to capitalize on diffusion models for video super-r
esolution which necessitates not only the preservation of visual appearance from
 low-resolution to high-resolution videos but also the temporal consistency acro
ss video frames. In this paper we propose a novel approach pursuing Spatial Adap
tation and Temporal Coherence (SATeCo) for video super-resolution. SATeCo pivots
 on learning spatial-temporal guidance from low-resolution videos to calibrate b
oth latent-space high-resolution video denoising and pixel-space video reconstru
ction. Technically SATeCo freezes all the parameters of the pre-trained UNet and
 VAE and only optimizes two deliberately-designed spatial feature adaptation (SF
A) and temporal feature alignment (TFA) modules in the decoder of UNet and VAE.
SFA modulates frame features via adaptively estimating affine parameters for eac
h pixel guaranteeing pixel-wise guidance for high-resolution frame synthesis. TF
A delves into feature interaction within a 3D local window (tubelet) through sel
f-attention and executes cross-attention between tubelet and its low-resolution
counterpart to guide temporal feature alignment. Extensive experiments conducted
 on the REDS4 and Vid4 datasets demonstrate the effectiveness of our approach.
****
Multiagent Multitraversal Multimodal Self-Driving: Open MARS Dataset
Yiming Li, Zhiheng Li, Nuo Chen, Moonjun Gong, Zonglin Lyu, Zehong Wang, Peili J
iang, Chen Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2024, pp. 22041-22051
Large-scale datasets have fueled recent advancements in AI-based autonomous vehi
cle research. However these datasets are usually collected from a single vehicle
's one-time pass of a certain location lacking multiagent interactions or repeat
ed traversals of the same place. Such information could lead to transformative e
nhancements in autonomous vehicles' perception prediction and planning capabilit
ies. To bridge this gap in collaboration with the self-driving company May Mobil
ity we present the MARS dataset which unifies scenarios that enable MultiAgent m
ultitraveRSal and multimodal autonomous vehicle research. More specifically MARS
 is collected with a fleet of autonomous vehicles driving within a certain geogr
aphical area. Each vehicle has its own route and different vehicles may appear a
t nearby locations. Each vehicle is equipped with a LiDAR and surround-view RGB
cameras. We curate two subsets in MARS: one facilitates collaborative driving wi
th multiple vehicles simultaneously present at the same location and the other e
nables memory retrospection through asynchronous traversals of the same location
 by multiple vehicles. We conduct experiments in place recognition and neural re
construction. More importantly MARS introduces new research opportunities and ch
allenges such as multitraversal 3D reconstruction multiagent perception and unsu
pervised object discovery. Our data and codes can be found at https://ai4ce.gith
ub.io/MARS/.
****
From Variance to Veracity: Unbundling and Mitigating Gradient Variance in Differ
entiable Bundle Adjustment Layers

Swaminathan Gurumurthy, Karnik Ram, Bingqing Chen, Zachary Manchester, Zico Kolter; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27507-27516

Various pose estimation and tracking problems in robotics can be decomposed into a correspondence estimation problem (often computed using a deep network) followed by a weighted least squares optimization problem to solve for the poses. Recent work has shown that coupling the two problems by iteratively refining one conditioned on the other's output yields SOTA results across domains. However training these models has proved challenging requiring a litany of tricks to stabilize and speed up training. In this work we take the visual odometry problem as an example and identify three plausible causes: (1) flow loss interference (2) linearization errors in the bundle adjustment (BA) layer and (3) dependence of weight gradients on the BA residual. We show how these issues result in noisy and higher variance gradients potentially leading to a slow down in training and instabilities. We then propose a simple solution to reduce the gradient variance by using the weights predicted by the network in the inner optimization loop to also weight the correspondence objective in the training problem. This helps the training objective 'focus' on the more important points thereby reducing the variance and mitigating the influence of outliers. We show that the resulting method leads to faster training and can be more flexibly trained in varying training setups without sacrificing performance. In particular we show 2-2.5x training speed ups over a baseline visual odometry model we modify.

****

Denoising Point Clouds in Latent Space via Graph Convolution and Invertible Neural Network

Aihua Mao, Biao Yan, Zijing Ma, Ying He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5768-5777

Point clouds frequently contain noise and outliers presenting obstacles for downstream applications. In this work we introduce a novel denoising method for point clouds. By leveraging the latent space we explicitly uncover noise components allowing for the extraction of a clean latent code. This in turn facilitates the restoration of clean points via inverse transformation. A key component in our network is a new multi-level graph convolution network for capturing rich geometric structural features at various scales from local to global. These features are then integrated into the invertible neural network which bijectively maps the latent space to guide the noise disentanglement process. Additionally we employ an invertible monotone operator to model the transformation process effectively enhancing the representation of integrated geometric features. This enhancement allows our network to precisely differentiate between noise factors and the intrinsic clean points in the latent code by projecting them onto separate channels. Both qualitative and quantitative evaluations demonstrate that our method outperforms state-of-the-art methods at various noise levels. The source code is available at https://github.com/yanbiao1/PD-LTS.

****

ADA-Track: End-to-End Multi-Camera 3D Multi-Object Tracking with Alternating Detection and Association

Shuxiao Ding, Lukas Schneider, Marius Cordts, Juergen Gall; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15184-15194

Many query-based approaches for 3D Multi-Object Tracking (MOT) adopt the tracking-by-attention paradigm utilizing track queries for identity-consistent detection and object queries for identity-agnostic track spawning. Tracking-by-attention however entangles detection and tracking queries in one embedding for both the detection and tracking task which is sub-optimal. Other approaches resemble the tracking-by-detection paradigm detecting objects using decoupled track and detection queries followed by a subsequent association. These methods however do not leverage synergies between the detection and association task. Combining the strengths of both paradigms we introduce ADA-Track a novel end-to-end framework for 3D MOT from multi-view cameras. We introduce a learnable data association module based on edge-augmented cross-attention leveraging appearance and geometric fe

atures. Furthermore we integrate this association module into the decoder layer of a DETR-based 3D detector enabling simultaneous DETR-like query-to-image cross-attention for detection and query-to-query cross-attention for data association. By stacking these decoder layers queries are refined for the detection and association task alternately effectively harnessing the task dependencies. We evaluate our method on the nuScenes dataset and demonstrate the advantage of our approach compared to the two previous paradigms. Code is available at https://github.com/dsx0511/ADA-Track.
****
HIR-Diff: Unsupervised Hyperspectral Image Restoration Via Improved Diffusion Models
Li Pang, Xiangyu Rui, Long Cui, Hongzhong Wang, Deyu Meng, Xiangyong Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3005-3014
Hyperspectral image (HSI) restoration aims at recovering clean images from degraded observations and plays a vital role in downstream tasks. Existing model-based methods have limitations in accurately modeling the complex image characteristics with handcraft priors and deep learning-based methods suffer from poor generalization ability. To alleviate these issues this paper proposes an unsupervised HSI restoration framework with pre-trained diffusion model (HIR-Diff) which restores the clean HSIs from the product of two low-rank components i.e. the reduced image and the coefficient matrix. Specifically the reduced image which has a low spectral dimension lies in the image field and can be inferred from our improved diffusion model where a new guidance function with total variation (TV) prior is designed to ensure that the reduced image can be well sampled. The coefficient matrix can be effectively pre-estimated based on singular value decomposition (SVD) and rank-revealing QR (RRQR) factorization. Furthermore a novel exponential noise schedule is proposed to accelerate the restoration process (about 5xacceleration for denoising) with little performance decrease. Extensive experimental results validate the superiority of our method in both performance and speed on a variety of HSI restoration tasks including HSI denoising noisy HSI super-resolution and noisy HSI inpainting. The code is available at https://github.com/LiPang/HIRDiff.
****
Mind The Edge: Refining Depth Edges in Sparsely-Supervised Monocular Depth Estimation
Lior Talker, Aviad Cohen, Erez Yosef, Alexandra Dana, Michael Dinerstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10606-10616
Monocular Depth Estimation (MDE) is a fundamental problem in computer vision with numerous applications. Recently LIDAR-supervised methods have achieved remarkable per-pixel depth accuracy in outdoor scenes. However significant errors are typically found in the proximity of depth discontinuities i.e. depth edges which often hinder the performance of depth-dependent applications that are sensitive to such inaccuracies e.g. novel view synthesis and augmented reality. Since direct supervision for the location of depth edges is typically unavailable in sparse LIDAR-based scenes encouraging the MDE model to produce correct depth edges is not straightforward. To the best of our knowledge this paper is the first attempt to address the depth edges issue for LIDAR-supervised scenes. In this work we propose to learn to detect the location of depth edges from densely-supervised synthetic data and use it to generate supervision for the depth edges in the MDE training. To quantitatively evaluate our approach and due to the lack of depth edges GT in LIDAR-based scenes we manually annotated subsets of the KITTI and the DDAD datasets with depth edges ground truth. We demonstrate significant gains in the accuracy of the depth edges with comparable per-pixel depth accuracy on several challenging datasets. Code and datasets are available at https://github.com/liortalker/MindTheEdge.
****
Attention-Driven Training-Free Efficiency Enhancement of Diffusion Models
Hongjie Wang, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K. Jha, Yuchen Liu;

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16080-16089
Diffusion models (DMs) have exhibited superior performance in generating high-quality and diverse images. However this exceptional performance comes at the cost of expensive generation process particularly due to the heavily used attention module in leading models. Existing works mainly adopt a retraining process to enhance DM efficiency. This is computationally expensive and not very scalable. To this end we introduce the Attention-driven Training-free Efficient Diffusion Model (AT-EDM) framework that leverages attention maps to perform run-time pruning of redundant tokens without the need for any retraining. Specifically for single-denoising-step pruning we develop a novel ranking algorithm Generalized Weighted Page Rank (G-WPR) to identify redundant tokens and a similarity-based recovery method to restore tokens for the convolution operation. In addition we propose a Denoising-Steps-Aware Pruning (DSAP) approach to adjust the pruning budget across different denoising timesteps for better generation quality. Extensive evaluations show that AT-EDM performs favorably against prior art in terms of efficiency (e.g. 38.8% FLOPs saving and up to 1.53x speed-up over Stable Diffusion XL) while maintaining nearly the same FID and CLIP scores as the full model.
****
CPR: Retrieval Augmented Generation for Copyright Protection
Aditya Golatkar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12374-12384
Retrieval Augmented Generation (RAG) is emerging as a flexible and robust technique to adapt models to private users data without training to handle credit attribution and to allow efficient machine unlearning at scale. However RAG techniques for image generation may lead to parts of the retrieved samples being copied in the model's output. To reduce risks of leaking private information contained in the retrieved set we introduce Copy-Protected generation with Retrieval (CPR) a new method for RAG with strong copyright protection guarantees in a mixed-private setting for diffusion models. CPR allows to condition the output of diffusion models on a set of retrieved images while also guaranteeing that unique identifiable information about those example is not exposed in the generated outputs. In particular it does so by sampling from a mixture of public (safe) distribution and private (user) distribution by merging their diffusion scores at inference. We prove that CPR satisfies Near Access Freeness (NAF) which bounds the amount of information an attacker may be able to extract from the generated images. We provide two algorithms for copyright protection CPR-KL and CPR-Choose. Unlike previously proposed rejection-sampling-based NAF methods our methods enable efficient copyright-protected sampling with a single run of backward diffusion. We show that our method can be applied to any pre-trained conditional diffusion model such as Stable Diffusion or unCLIP. In particular we empirically show that applying CPR on top of un- CLIP improves quality and text-to-image alignment of the generated results (81.4 to 83.17 on TIFA benchmark) while enabling credit attribution copy-right protection and deterministic constant time unlearning.
****
FreeDrag: Feature Dragging for Reliable Point-based Image Editing
Pengyang Ling, Lin Chen, Pan Zhang, Huaian Chen, Yi Jin, Jinjin Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6860-6870
To serve the intricate and varied demands of image editing precise and flexible manipulation in image content is indispensable. Recently Drag-based editing methods have gained impressive performance. However these methods predominantly center on point dragging resulting in two noteworthy drawbacks namely "miss tracking" where difficulties arise in accurately tracking the predetermined handle points and "ambiguous tracking" where tracked points are potentially positioned in wrong regions that closely resemble the handle points. To address the above issues we propose FreeDrag a feature dragging methodology designed to free the burden on point tracking. The FreeDrag incorporates two key designs i.e. template feature via adaptive updating and line search with backtracking the former improves t

he stability against drastic content change by elaborately controlling the featu
re updating scale after each dragging while the latter alleviates the misguidanc
e from similar points by actively restricting the search area in a line. These t
wo technologies together contribute to a more stable semantic dragging with high
er efficiency. Comprehensive experimental results substantiate that our approach
 significantly outperforms pre-existing methodologies offering reliable point-ba
sed editing even in various complex scenarios.
****
Image-Text Co-Decomposition for Text-Supervised Semantic Segmentation
Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-
Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, Yen-Yu Lin; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26794-26
803
This paper addresses text-supervised semantic segmentation aiming to learn a mod
el capable of segmenting arbitrary visual concepts within images by using only i
mage-text pairs without dense annotations. Existing methods have demonstrated th
at contrastive learning on image-text pairs effectively aligns visual segments w
ith the meanings of texts. We notice that there is a discrepancy between text al
ignment and semantic segmentation: A text often consists of multiple semantic co
ncepts whereas semantic segmentation strives to create semantically homogeneous
segments. To address this issue we propose a novel framework Image-Text Co-Decom
position (CoDe) where the paired image and text are jointly decomposed into a se
t of image regions and a set of word segments respectively and contrastive learn
ing is developed to enforce region-word alignment. To work with a vision-languag
e model we present a prompt learning mechanism that derives an extra representat
ion to highlight an image segment or a word segment of interest with which more
effective features can be extracted from that segment. Comprehensive experimenta
l results demonstrate that our method performs favorably against existing text-s
upervised semantic segmentation methods on six benchmark datasets.
****
Orchestrate Latent Expertise: Advancing Online Continual Learning with Multi-Lev
el Supervision and Reverse Self-Distillation
Hongwei Yan, Liyuan Wang, Kaisheng Ma, Yi Zhong; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23670-23680
To accommodate real-world dynamics artificial intelligence systems need to cope
with sequentially arriving content in an online manner. Beyond regular Continual
 Learning (CL) attempting to address catastrophic forgetting with offline traini
ng of each task Online Continual Learning (OCL) is a more challenging yet realis
tic setting that performs CL in a one-pass data stream. Current OCL methods prim
arily rely on memory replay of old training samples. However a notable gap from
CL to OCL stems from the additional overfitting-underfitting dilemma associated
with the use of rehearsal buffers: the inadequate learning of new training sampl
es (underfitting) and the repeated learning of a few old training samples (overf
itting). To this end we introduce a novel approach Multi-level Online Sequential
 Experts (MOSE) which cultivates the model as stacked sub-experts integrating mu
lti-level supervision and reverse self-distillation. Supervision signals across
multiple stages facilitate appropriate convergence of the new task while gatheri
ng various strengths from experts by knowledge distillation mitigates the perfor
mance decline of old tasks. MOSE demonstrates remarkable efficacy in learning ne
w samples and preserving past knowledge through multi-level experts thereby sign
ificantly advancing OCL performance over state-of-the-art baselines (e.g. up to
7.3% on Split CIFAR-100 and 6.1% on Split Tiny-ImageNet).
****
Vision-and-Language Navigation via Causal Learning
Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, Qijun Chen; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 13139-13150
In the pursuit of robust and generalizable environment perception and language u
nderstanding the ubiquitous challenge of dataset bias continues to plague vision
-and-language navigation (VLN) agents hindering their performance in unseen envi

ronments. This paper introduces the generalized cross-modal causal transformer (
GOAT) a pioneering solution rooted in the paradigm of causal inference. By delvi
ng into both observable and unobservable confounders within vision language and
history we propose the back-door and front-door adjustment causal learning (BACL
 and FACL) modules to promote unbiased learning by comprehensively mitigating po
tential spurious correlations. Additionally to capture global confounder feature
s we propose a cross-modal feature pooling (CFP) module supervised by contrastiv
e learning which is also shown to be effective in improving cross-modal represen
tations during pre-training. Extensive experiments across multiple VLN datasets
(R2R REVERIE RxR and SOON) underscore the superiority of our proposed method ove
r previous state-of-the-art approaches. Code is available at https://github.com/
CrystalSixone/VLN-GOAT.
****
Mitigating Object Dependencies: Improving Point Cloud Self-Supervised Learning t
hrough Object Exchange
Yanhao Wu, Tong Zhang, Wei Ke, Congpei Qiu, Sabine Süsstrunk, Mathieu Salzmann;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2024, pp. 23052-23061
In the realm of point cloud scene understanding particularly in indoor scenes ob
jects are arranged following human habits resulting in objects of certain semant
ics being closely positioned and displaying notable inter-object correlations. T
his can create a tendency for neural networks to exploit these strong dependenci
es bypassing the individual object patterns. To address this challenge we introd
uce a novel self-supervised learning (SSL) strategy. Our approach leverages both
 object patterns and contextual cues to produce robust features. It begins with
the formulation of an object-exchanging strategy where pairs of objects with com
parable sizes are exchanged across different scenes effectively disentangling th
e strong contextual dependencies. Subsequently we introduce a context-aware feat
ure learning strategy which encodes object patterns without relying on their spe
cific context by aggregating object features across various scenes. Our extensiv
e experiments demonstrate the superiority of our method over existing SSL techni
ques further showing its better robustness to environmental changes. Moreover we
 showcase the applicability of our approach by transferring pre-trained models t
o diverse point cloud datasets.
****
Confronting Ambiguity in 6D Object Pose Estimation via Score-Based Diffusion on
SE(3)
Tsu-Ching Hsiao, Hao-Wei Chen, Hsuan-Kung Yang, Chun-Yi Lee; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 352-362
Addressing pose ambiguity in 6D object pose estimation from single RGB images pr
esents a significant challenge particularly due to object symmetries or occlusio
ns. In response we introduce a novel score-based diffusion method applied to the
 SE(3) group marking the first application of diffusion models to SE(3) within t
he image domain specifically tailored for pose estimation tasks. Extensive evalu
ations demonstrate the method's efficacy in handling pose ambiguity mitigating p
erspective-induced ambiguity and showcasing the robustness of our surrogate Stei
n score formulation on SE(3). This formulation not only improves the convergence
 of denoising process but also enhances computational efficiency. Thus we pionee
r a promising strategy for 6D object pose estimation.
****
Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models
Daniel Geng, Inbum Park, Andrew Owens; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24154-24163
We address the problem of synthesizing multi-view optical illusions: images that
 change appearance upon a transformation such as a flip or rotation. We propose
a simple zero-shot method for obtaining these illusions from off-the-shelf text-
to-image diffusion models. During the reverse diffusion process we estimate the
noise from different views of a noisy image and then combine these noise estimat
es together and denoise the image. A theoretical analysis suggests that this met

hod works precisely for views that can be written as orthogonal transformations of which permutations are a subset. This leads to the idea of a visual anagram ---an image that changes appearance under some rearrangement of pixels. This includes rotations and flips but also more exotic pixel permutations such as a jigsaw rearrangement. Our approach also naturally extends to illusions with more than two views. We provide both qualitative and quantitative results demonstrating the effectiveness and flexibility of our method. Please see our project webpage for additional visualizations and results: https://dangeng.github.io/visual_anagrams/
****
Unveiling Parts Beyond Objects: Towards Finer-Granularity Referring Expression Segmentation
Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, Jing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12998-13008
Referring expression segmentation (RES) aims at segmenting the foreground masks of the entities that match the descriptive natural language expression. Previous datasets and methods for classic RES task heavily rely on the prior assumption that one expression must refer to object-level targets. In this paper we take a step further to finer-grained part-level RES task. To promote the object-level RES task towards finer-grained vision-language understanding we put forward a new multi-granularity referring expression segmentation (MRES) task and construct an evaluation benchmark called RefCOCOm by manual annotations. By employing our automatic model-assisted data engine we build the largest visual grounding dataset namely MRES-32M which comprises over 32.2M high-quality masks and captions on the provided 1M images. Besides a simple yet strong model named UniRES is designed to accomplish the unified object-level and part-level grounding task. Extensive experiments on our RefCOCOm for MRES and three datasets (i.e. RefCOCO(+/g)) for classic RES task demonstrate the superiority of our method over previous state-of-the-art methods. To foster future research into fine-grained visual grounding our benchmark RefCOCOm the MRES-32M dataset and model UniRES will be publicly available at https://github.com/Rubics-Xuan/MRES.
****
DiffInDScene: Diffusion-based High-Quality 3D Indoor Scene Generation
Xiaoliang Ju, Zhaoyang Huang, Yijin Li, Guofeng Zhang, Yu Qiao, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4526-4535
We present DiffInDScene a novel framework for tackling the problem of high-quality 3D indoor scene generation which is challenging due to the complexity and diversity of the indoor scene geometry. Although diffusion-based generative models have previously demonstrated impressive performance in image generation and object-level 3D generation they have not yet been applied to room-level 3D generation due to their computationally intensive costs. In DiffInDScene we propose a cascaded 3D diffusion pipeline that is efficient and possesses strong generative performance for Truncated Signed Distance Function (TSDF). The whole pipeline is designed to run on a sparse occupancy space in a coarse-to-fine fashion. Inspired by KinectFusion's incremental alignment and fusion of local TSDF volumes we propose a diffusion-based SDF fusion approach that iteratively diffuses and fuses local TSDF volumes facilitating the generation of an entire room environment. The generated results demonstrate that our work is capable to achieve high-quality room generation directly in three-dimensional space starting from scratch. In addition to the scene generation the final part of DiffInDScene can be used as a post-processing module to refine the 3D reconstruction results from multi-view stereo. According to the user study the mesh quality generated by our DiffInDScene can even outperform the ground truth mesh provided by ScanNet.
****
MAPSeg: Unified Unsupervised Domain Adaptation for Heterogeneous Medical Image Segmentation Based on 3D Masked Autoencoding and Pseudo-Labeling
Xuzhe Zhang, Yuhao Wu, Elsa Angelini, Ang Li, Jia Guo, Jerod M. Rasmussen, Thomas G. O'Connor, Pathik D. Wadhwa, Andrea Parolin Jackowski, Hai Li, Jonathan Posn

er, Andrew F. Laine, Yun Wang; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2024, pp. 5851-5862
Robust segmentation is critical for deriving quantitative measures from large-sc
ale multi-center and longitudinal medical scans. Manually annotating medical sca
ns however is expensive and labor-intensive and may not always be available in e
very domain. Unsupervised domain adaptation (UDA) is a well-studied technique th
at alleviates this label-scarcity problem by leveraging available labels from an
other domain. In this study we introduce Masked Autoencoding and Pseudo-Labeling
 Segmentation (MAPSeg) a unified UDA framework with great versatility and superi
or performance for heterogeneous and volumetric medical image segmentation. To t
he best of our knowledge this is the first study that systematically reviews and
 develops a framework to tackle four different domain shifts in medical image se
gmentation. More importantly MAPSeg is the first framework that can be applied t
o centralized federated and test-time UDA while maintaining comparable performan
ce. We compare MAPSeg with previous state-of-the-art methods on a private infant
 brain MRI dataset and a public cardiac CT-MRI dataset and MAPSeg outperforms ot
hers by a large margin (10.5 Dice improvement on the private MRI dataset and 5.7
 on the public CT-MRI dataset). MAPSeg poses great practical value and can be ap
plied to real-world problems. GitHub: https://github.com/XuzheZ/MAPSeg/.
****
Leveraging Predicate and Triplet Learning for Scene Graph Generation
Jiankai Li, Yunhong Wang, Xiefan Guo, Ruijie Yang, Weixin Li; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 28369-28379
Scene Graph Generation (SGG) aims to identify entities and predict the relations
hip triplets <subject predicate object> in visual scenes. Given the prevalence o
f large visual variations of subject-object pairs even in the same predicate it
can be quite challenging to model and refine predicate representations directly
across such pairs which is however a common strategy adopted by most existing SG
G methods. We observe that visual variations within the identical triplet are re
latively small and certain relation cues are shared in the same type of triplet
which can potentially facilitate the relation learning in SGG. Moreover for the
long-tail problem widely studied in SGG task it is also crucial to deal with the
 limited types and quantity of triplets in tail predicates. Accordingly in this
paper we propose a Dual-granularity Relation Modeling (DRM) network to leverage
fine-grained triplet cues besides the coarse-grained predicate ones. DRM utilize
s contexts and semantics of predicate and triplet with Dual-granularity Constrai
nts generating compact and balanced representations from two perspectives to fac
ilitate relation recognition. Furthermore a Dual-granularity Knowledge Transfer
(DKT) strategy is introduced to transfer variation from head predicates/triplets
 to tail ones aiming to enrich the pattern diversity of tail classes to alleviat
e the long-tail problem. Extensive experiments demonstrate the effectiveness of
our method which establishes new state-of-the-art performance on Visual Genome O
pen Image and GQA datasets. Our code is available at https://github.com/jkli1998
/DRM.
****
DaReNeRF: Direction-aware Representation for Dynamic Scenes
Ange Lou, Benjamin Planche, Zhongpai Gao, Yamin Li, Tianyu Luan, Hao Ding, Terre
nce Chen, Jack Noble, Ziyan Wu; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2024, pp. 5031-5042
Addressing the intricate challenge of modeling and re-rendering dynamic scenes m
ost recent approaches have sought to simplify these complexities using plane-bas
ed explicit representations overcoming the slow training time issues associated
with methods like Neural Radiance Fields (NeRF) and implicit representations. Ho
wever the straightforward decomposition of 4D dynamic scenes into multiple 2D pl
ane-based representations proves insufficient for re-rendering high-fidelity sce
nes with complex motions. In response we present a novel direction-aware represe
ntation (DaRe) approach that captures scene dynamics from six different directio
ns. This learned representation undergoes an inverse dual-tree complex wavelet t
ransformation (DTCWT) to recover plane-based information. DaReNeRF computes feat

ures for each space-time point by fusing vectors from these recovered planes. Co
mbining DaReNeRF with a tiny MLP for color regression and leveraging volume rend
ering in training yield state-of-the-art performance in novel view synthesis for
 complex dynamic scenes. Notably to address redundancy introduced by the six rea
l and six imaginary direction-aware wavelet coefficients we introduce a trainabl
e masking approach mitigating storage issues without significant performance dec
line. Moreover DaReNeRF maintains a 2x reduction in training time compared to pr
ior art while delivering superior performance.
****
SfmCAD: Unsupervised CAD Reconstruction by Learning Sketch-based Feature Modelin
g Operations
Pu Li, Jianwei Guo, Huibin Li, Bedrich Benes, Dong-Ming Yan; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 4671-4680
This paper introduces SfmCAD a novel unsupervised network that reconstructs 3D s
hapes by learning the Sketch-based Feature Modeling operations commonly used in
modern CAD workflows. Given a 3D shape represented as voxels SfmCAD learns a neu
ral-typed sketch+path parameterized representation including 2D sketches of feat
ure primitives and their 3D sweeping paths without supervision for inferring fea
ture-based CAD programs. SfmCAD employs 2D sketches for local detail representat
ion and 3D paths to capture the overall structure achieving a clear separation b
etween shape details and structure. This conversion into parametric forms enable
s users to seamlessly adjust the shape's geometric and structural features thus
enhancing interpretability and user control. We demonstrate the effectiveness of
 our method by applying SfmCAD to many different types of objects such as CAD pa
rts ShapeNet objects and tree shapes. Extensive comparisons show that SfmCAD pro
duces compact and faithful 3D reconstructions with superior quality compared to
alternatives. The code is released at https://github.com/BunnySoCrazy/SfmCAD.
****
CoDi-2: In-Context Interleaved and Interactive Any-to-Any Generation
Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, Mohit Bansal;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2024, pp. 27425-27434
We present CoDi-2 a Multimodal Large Language Model (MLLM) for learning in-conte
xt interleaved multimodal representations. By aligning modalities with language
for both encoding and generation CoDi-2 empowers Large Language Models (LLMs) to
 understand modality-interleaved instructions and in-context examples and autore
gressively generate grounded and coherent multimodal outputs in an any-to-any in
put-output modality paradigm. To train CoDi-2 we build a large-scale generation
dataset encompassing in-context multimodal instructions across text vision and a
udio. CoDi-2 demonstrates a wide range of zero-shot and few-shot capabilities fo
r tasks like editing exemplar learning composition reasoning etc. CoDi-2 surpass
es previous domain-specific models on tasks such as subject-driven image generat
ion vision transformation and audio editing and showcases a significant advancem
ent for integrating diverse multimodal tasks with sequential generation.
****
Tuning Stable Rank Shrinkage: Aiming at the Overlooked Structural Risk in Fine-t
uning
Sicong Shen, Yang Zhou, Bingzheng Wei, Eric I-Chao Chang, Yan Xu; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 28474-28484
Existing fine-tuning methods for computer vision tasks primarily focus on re-wei
ghting the knowledge learned from the source domain during pre-training. They ai
m to retain beneficial knowledge for the target domain while suppressing unfavor
able knowledge. During the pre-training and fine-tuning stages there is a notabl
e disparity in the data scale. Consequently it is theoretically necessary to emp
loy a model with reduced complexity to mitigate the potential structural risk. H
owever our empirical investigation in this paper reveals that models fine-tuned
using existing methods still manifest a high level of model complexity inherited
 from the pre-training stage leading to a suboptimal stability and generalizatio

n ability. This phenomenon indicates an issue that has been overlooked in fine-tuning: Structural Risk Minimization. To address this issue caused by data scale disparity during the fine-tuning stage we propose a simple yet effective approach called Tuning Stable Rank Shrinkage (TSRS). TSRS mitigates the structural risk during the fine-tuning stage by constraining the noise sensitivity of the target model based on stable rank theories. Through extensive experiments we demonstrate that incorporating TSRS into fine-tuning methods leads to improved generalization ability on various tasks regardless of whether the neural networks are based on convolution or transformer architectures. Additionally empirical analysis reveals that TSRS enhances the robustness convexity and smoothness of the loss landscapes in fine-tuned models. Code is available at https://github.com/WitGotFlg/TSRS.
****
Differentiable Display Photometric Stereo
Seokjun Choi, Seungwoo Yoon, Giljoo Nam, Seungyong Lee, Seung-Hwan Baek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11831-11840
Photometric stereo leverages variations in illumination conditions to reconstruct surface normals. Display photometric stereo which employs a conventional monitor as an illumination source has the potential to overcome limitations often encountered in bulky and difficult-to-use conventional setups. In this paper we present differentiable display photometric stereo (DDPS) addressing an often overlooked challenge in display photometric stereo: the design of display patterns. Departing from using heuristic display patterns DDPS learns the display patterns that yield accurate normal reconstruction for a target system in an end-to-end manner. To this end we propose a differentiable framework that couples basis-illumination image formation with analytic photometric-stereo reconstruction. The differentiable framework facilitates the effective learning of display patterns via auto-differentiation. Also for training supervision we propose to use 3D printing for creating a real-world training dataset enabling accurate reconstruction on the target real-world setup. Finally we exploit that conventional LCD monitors emit polarized light which allows for the optical separation of diffuse and specular reflections when combined with a polarization camera leading to accurate normal reconstruction. Extensive evaluation of DDPS shows improved normal-reconstruction accuracy compared to heuristic patterns and demonstrates compelling properties such as robustness to pattern initialization calibration errors and simplifications in image formation and reconstruction.
****
In-distribution Public Data Synthesis with Diffusion Models for Differentially Private Image Classification
Jinseong Park, Yujin Choi, Jaewook Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12236-12246
To alleviate the utility degradation of deep learning image classification with differential privacy (DP) employing extra public data or pre-trained models has been widely explored. Recently the use of in-distribution public data has been investigated where tiny subsets of datasets are released publicly. In this paper we investigate a framework that leverages recent diffusion models to amplify the information of public data. Subsequently we identify data diversity and generalization gap between public and private data as critical factors addressing the limited public data. While assuming 4% of training data as public our method achieves 85.48% on CIFAR-10 with a privacy budget of ?=2 without employing extra public data for training.
****
Learning Degradation-unaware Representation with Prior-based Latent Transformations for Blind Face Restoration
Lianxin Xie, Csbingbing Zheng, Wen Xue, Le Jiang, Cheng Liu, Si Wu, Hau San Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9120-9129
Blind face restoration focuses on restoring high-fidelity details from images subjected to complex and unknown degradations while preserving identity informatio

n. In this paper we present a Prior-based Latent Transformation approach (PLTrans) which is specifically designed to learn a degradation-unaware representation thereby allowing the restoration network to effectively generalize to real-world degradation. Toward this end PLTrans learns a degradation-unaware query via a latent diffusion-based regularization module. Furthermore conditioned on the features of a degraded face image a latent dictionary that captures the priors of HQ face images is leveraged to refine the features by mapping the top-d nearest elements. The refined version will be used to build key and value for the cross-attention computation which is tailored to each degraded image and exhibits reduced sensitivity to different degradation factors. Conditioned on the resulting representation we train a decoding network that synthesizes face images with authentic details and identity preservation. Through extensive experiments we verify the effectiveness of the design elements and demonstrate the generalization ability of our proposed approach for both synthetic and unknown degradations. We finally demonstrate the applicability of PLTrans in other vision tasks.
****
LSK3DNet: Towards Effective and Efficient 3D Perception with Large Sparse Kernels
Tuo Feng, Wenguan Wang, Fan Ma, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14916-14927
Autonomous systems need to process large-scale sparse and irregular point clouds with limited compute resources. Consequently it is essential to develop LiDAR perception methods that are both efficient and effective. Although naively enlarging 3D kernel size can enhance performance it will also lead to a cubically-increasing overhead. Therefore it is crucial to develop streamlined 3D large kernel designs that eliminate redundant weights and work effectively with larger kernels. In this paper we propose an efficient and effective Large Sparse Kernel 3D Neural Network (LSK3DNet) that leverages dynamic pruning to amplify the 3D kernel size. Our method comprises two core components: Spatial-wise Dynamic Sparsity (SDS) and Channel-wise Weight Selection (CWS). SDS dynamically prunes and regrows volumetric weights from the beginning to learn a large sparse 3D kernel. It not only boosts performance but also significantly reduces model size and computational cost. Moreover CWS selects the most important channels for 3D convolution during training and subsequently prunes the redundant channels to accelerate inference for 3D vision tasks. We demonstrate the effectiveness of LSK3DNet on three benchmark datasets and five tracks compared with classical models and large kernel designs. Notably LSK3DNet achieves the state-of-the-art performance on SemanticKITTI (i.e. 75.6% on single-scan and 63.4% on multi-scan) with roughly 40% model size reduction and 60% computing operations reduction compared to the naive large 3D kernel model.
****
Faces that Speak: Jointly Synthesising Talking Face and Speech from Text
Youngjoon Jang, Ji-Hoon Kim, Junseok Ahn, Doyeop Kwak, Hong-Sun Yang, Yoon-Cheol Ju, Il-Hwan Kim, Byeong-Yeol Kim, Joon Son Chung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8818-8828
The goal of this work is to simultaneously generate natural talking faces and speech outputs from text. We achieve this by integrating Talking Face Generation (TFG) and Text-to-Speech (TTS) systems into a unified framework. We address the main challenges of each task: (1) generating a range of head poses representative of real-world scenarios and (2) ensuring voice consistency despite variations in facial motion for the same identity. To tackle these issues we introduce a motion sampler based on conditional flow matching which is capable of high-quality motion code generation in an efficient way. Moreover we introduce a novel conditioning method for the TTS system which utilises motion-removed features from the TFG model to yield uniform speech outputs. Our extensive experiments demonstrate that our method effectively creates natural-looking talking faces and speech that accurately match the input text. To our knowledge this is the first effort to build a multimodal synthesis system that can generalise to unseen identities.
****
Diversified and Personalized Multi-rater Medical Image Segmentation

Yicheng Wu, Xiangde Luo, Zhe Xu, Xiaoqing Guo, Lie Ju, Zongyuan Ge, Wenjun Liao, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11470-11479
Annotation ambiguity due to inherent data uncertainties such as blurred boundaries in medical scans and different observer expertise and preferences has become a major obstacle for training deep-learning based medical image segmentation models. To address it the common practice is to gather multiple annotations from different experts leading to the setting of multi-rater medical image segmentation. Existing works aim to either merge different annotations into the "groundtruth" that is often unattainable in numerous medical contexts or generate diverse results or produce personalized results corresponding to individual expert raters. Here we bring up a more ambitious goal for multi-rater medical image segmentation i.e. obtaining both diversified and personalized results. Specifically we propose a two-stage framework named D-Persona (first Diversification and then Personalization). In Stage I we exploit multiple given annotations to train a Probabilistic U-Net model with a bound-constrained loss to improve the prediction diversity. In this way a common latent space is constructed in Stage I where different latent codes denote diversified expert opinions. Then in Stage II we design multiple attention-based projection heads to adaptively query the corresponding expert prompts from the shared latent space and then perform the personalized medical image segmentation. We evaluated the proposed model on our in-house Nasopharyngeal Carcinoma dataset and the public lung nodule dataset (i.e. LIDC-IDRI). Extensive experiments demonstrated our D-Persona can provide diversified and personalized results at the same time achieving new SOTA performance for multi-rater medical image segmentation. Our code will be released at https://github.com/ycwu1997/D-Persona.
****
Towards Automatic Power Battery Detection: New Challenge Benchmark Dataset and Baseline
Xiaoqi Zhao, Youwei Pang, Zhenyu Chen, Qian Yu, Lihe Zhang, Hanqi Liu, Jiaming Zuo, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22020-22029
We conduct a comprehensive study on a new task named power battery detection (PBD) which aims to localize the dense cathode and anode plates endpoints from X-ray images to evaluate the quality of power batteries. Existing manufacturers usually rely on human eye observation to complete PBD which makes it difficult to balance the accuracy and efficiency of detection. To address this issue and drive more attention into this meaningful task we first elaborately collect a dataset called X-ray PBD which has 1500 diverse X-ray images selected from thousands of power batteries of 5 manufacturers with 7 different visual interference. Then we propose a novel segmentation-based solution for PBD termed multi-dimensional collaborative network (MDCNet). With the help of line and counting predictors the representation of the point segmentation branch can be improved at both semantic and detail aspects. Besides we design an effective distance-adaptive mask generation strategy which can alleviate the visual challenge caused by the inconsistent distribution density of plates to provide MDCNet with stable supervision. Without any bells and whistles our segmentation-based MDCNet consistently outperforms various other corner detection crowd counting and general/tiny object detection-based solutions making it a strong baseline that can help facilitate future research in PBD. Finally we share some potential difficulties and works for future researches. The source code and datasets will be publicly available at \href https://github.com/Xiaoqi-Zhao-DLUT/X-ray-PBD X-ray PBD .
****
AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection
Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, Gaurav Bharaj; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27102-27112
With the rapid growth in deepfake video content we require improved and generalizable methods to detect them. Most existing detection methods either use uni-modal cues or rely on supervised training to capture the dissonance between the aud

io and visual modalities. While the former disregards the audio-visual correspon dences entirely the latter predominantly focuses on discerning audio-visual cues within the training corpus thereby potentially overlooking correspondences that can help detect unseen deepfakes. We present Audio-Visual Feature Fusion (AVFF) a two-stage cross-modal learning method that explicitly captures the correspond ence between the audio and visual modalities for improved deepfake detection. Th e first stage pursues representation learning via self-supervision on real video s to capture the intrinsic audio-visual correspondences. To extract rich cross-m odal representations we use contrastive learning and autoencoding objectives and introduce a novel audio-visual complementary masking and feature fusion strateg y. The learned representations are tuned in the second stage where deepfake clas sification is pursued via supervised learning on both real and fake videos. Exte nsive experiments and analysis suggest that our novel representation learning pa radigm is highly discriminative in nature. We report 98.6% accuracy and 99.1% AU C on the FakeAVCeleb dataset outperforming the current audio-visual state-of-the -art by 14.9% and 9.9% respectively.
****
Discover and Mitigate Multiple Biased Subgroups in Image Classifiers
Zeliang Zhang, Mingqian Feng, Zhiheng Li, Chenliang Xu; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1090 6-10915
Machine learning models can perform well on in-distribution data but often fail on biased subgroups that are underrepresented in the training data hindering the robustness of models for reliable applications. Such subgroups are typically un known due to the absence of subgroup labels. Discovering biased subgroups is the key to understanding models' failure modes and further improving models' robust ness. Most previous works of subgroup discovery make an implicit assumption that models only underperform on a single biased subgroup which does not hold on in- the-wild data where multiple biased subgroups exist. In this work we propose Dec omposition Interpretation and Mitigation (DIM) a novel method to address a more challenging but also more practical problem of discovering multiple biased subgr oups in image classifiers. Our approach decomposes the image features into multi ple components that represent multiple subgroups. This decomposition is achieved via a bilinear dimension reduction method Partial Least Square (PLS) guided by useful supervision from the image classifier. We further interpret the semantic meaning of each subgroup component by generating natural language descriptions u sing vision-language foundation models. Finally DIM mitigates multiple biased su bgroups simultaneously via two strategies including the data- and model-centric strategies. Extensive experiments on CIFAR-100 and Breeds datasets demonstrate t he effectiveness of DIM in discovering and mitigating multiple biased subgroups. Furthermore DIM uncovers the failure modes of the classifier on Hard ImageNet s howcasing its broader applicability to understanding model bias in image classif iers.
****
DiffusionRegPose: Enhancing Multi-Person Pose Estimation using a Diffusion-Based End-to-End Regression Approach
Dayi Tan, Hansheng Chen, Wei Tian, Lu Xiong; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2230-2239
This paper presents the DiffusionRegPose a novel approach to multi-person pose e stimation that converts a one-stage end-to-end keypoint regression model into a diffusion-based sampling process. Existing one-stage deterministic regression me thods though efficient are often prone to missed or false detections in crowded or occluded scenes due to their inability to reason pose ambiguity. To address t hese challenges we handle ambiguous poses in a generative fashion i.e. sampling from the image-conditioned pose distributions characterized by a diffusion proba bilistic model. Specifically with initial pose tokens extracted from the image n oisy pose candidates are progressively refined by interacting with the initial t okens via attention layers. Extensive evaluations on the COCO and CrowdPose data sets show that DiffusionRegPose clearly improves the pose accuracy in crowded sc enarios as evidenced by a notable 4.0 AP increase in the AP_H metric on the Crow

dPose dataset. This demonstrates the model's potential for robust and precise human pose estimation in real-world applications.
****
Memory-Scalable and Simplified Functional Map Learning
Robin Magnet, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4041-4050
Deep functional maps have emerged in recent years as a prominent learning-based framework for non-rigid shape matching problems. While early methods in this domain only focused on learning in the functional domain the latest techniques have demonstrated that by promoting consistency between functional and pointwise maps leads to significant improvements in accuracy. Unfortunately existing approaches rely heavily on the computation of large dense matrices arising from soft pointwise maps which compromises their efficiency and scalability. To address this limitation we introduce a novel memory-scalable and efficient functional map learning pipeline. By leveraging the specific structure of functional maps we offer the possibility to achieve identical results without ever storing the pointwise map in memory. Furthermore based on the same approach we present a differentiable map refinement layer adapted from an existing axiomatic refinement algorithm. Unlike many functional map learning methods which use this algorithm at a post-processing step ours can be easily used at train time enabling to enforce consistency between the refined and initial versions of the map. Our resulting approach is both simpler more efficient and more numerically stable by avoiding differentiation through a linear system while achieving close to state-of-the-art results in challenging scenarios.
****
X-MIC: Cross-Modal Instance Conditioning for Egocentric Action Generalization
Anna Kukleva, Fadime Sener, Edoardo Remelli, Bugra Tekin, Eric Sauser, Bernt Schiele, Shugao Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26364-26373
Lately there has been growing interest in adapting vision-language models (VLMs) to image and third-person video classification due to their success in zero-shot recognition. However the adaptation of these models to egocentric videos has been largely unexplored. To address this gap we propose a simple yet effective cross-modal adaptation framework which we call X-MIC. Using a video adapter our pipeline learns to align frozen text embeddings to each egocentric video directly in the shared embedding space. Our novel adapter architecture retains and improves generalization of the pre-trained VLMs by disentangling learnable temporal modeling and frozen visual encoder. This results in an enhanced alignment of text embeddings to each egocentric video leading to a significant improvement in cross-dataset generalization. We evaluate our approach on the Epic-Kitchens Ego4D and EGTEA datasets for fine-grained cross-dataset action generalization demonstrating the effectiveness of our method.
****
ExMap: Leveraging Explainability Heatmaps for Unsupervised Group Robustness to Spurious Correlations
Rwiddhi Chakraborty, Adrian Sletten, Michael C. Kampffmeyer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12017-12026
Group robustness strategies aim to mitigate learned biases in deep learning models that arise from spurious correlations present in their training datasets. However most existing methods rely on the access to the label distribution of the groups which is time-consuming and expensive to obtain. As a result unsupervised group robustness strategies are sought. Based on the insight that a trained model's classification strategies can be inferred accurately based on explainability heatmaps we introduce ExMap an unsupervised two stage mechanism designed to enhance group robustness in traditional classifiers. ExMap utilizes a clustering module to infer pseudo-labels based on a model's explainability heatmaps which are then used during training in lieu of actual labels. Our empirical studies validate the efficacy of ExMap - We demonstrate that it bridges the per- formance gap with its supervised counterparts and outperforms existing partially supervised

and unsupervised methods. Additionally ExMap can be seamlessly integrated with e xisting group robustness learning strategies. Finally we demonstrate its potenti al in tackling the emerging issue of multiple shortcut mitigation
****
Gaussian Head Avatar: Ultra High-fidelity Head Avatar via Dynamic Gaussians
Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, Yebi n Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2024, pp. 1931-1941
Creating high-fidelity 3D head avatars has always been a research hotspot but th ere remains a great challenge under lightweight sparse view setups. In this pape r we propose Gaussian Head Avatar represented by controllable 3D Gaussians for h igh-fidelity head avatar modeling. We optimize the neutral 3D Gaussians and a fu lly learned MLP-based deformation field to capture complex expressions. The two parts benefit each other thereby our method can model fine-grained dynamic detai ls while ensuring expression accuracy. Furthermore we devise a well-designed geo metry-guided initialization strategy based on implicit SDF and Deep Marching Tet rahedra for the stability and convergence of the training procedure. Experiments show our approach outperforms other state-of-the-art sparse-view methods achiev ing ultra high-fidelity rendering quality at 2K resolution even under exaggerate d expressions. Project page: https://yuelangx.github.io/gaussianheadavatar.
****
Stratified Avatar Generation from Sparse Observations
Han Feng, Wenchao Ma, Quankai Gao, Xianwei Zheng, Nan Xue, Huijuan Xu; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 153-163
Estimating 3D full-body avatars from AR/VR devices is essential for creating imm ersive experiences in AR/VR applications. This task is challenging due to the li mited input from Head Mounted Devices which capture only sparse observations fro m the head and hands. Predicting the full-body avatars particularly the lower bo dy from these sparse observations presents significant difficulties. In this pap er we are inspired by the inherent property of the kinematic tree defined in the Skinned Multi-Person Linear (SMPL) model where the upper body and lower body sh are only one common ancestor node bringing the potential of decoupled reconstruc tion. We propose a stratified approach to decouple the conventional full-body av atar reconstruction pipeline into two stages with the reconstruction of the uppe r body first and a subsequent reconstruction of the lower body conditioned on th e previous stage. To implement this straightforward idea we leverage the latent diffusion model as a powerful probabilistic generator and train it to follow the latent distribution of decoupled motions explored by a VQ-VAE encoder-decoder m odel. Extensive experiments on AMASS mocap dataset demonstrate our state-of-the- art performance in the reconstruction of full-body motions.
****
Learning to Segment Referred Objects from Narrated Egocentric Videos
Yuhan Shen, Huiyu Wang, Xitong Yang, Matt Feiszli, Ehsan Elhamifar, Lorenzo Torr esani, Effrosyni Mavroudi; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2024, pp. 14510-14520
Egocentric videos provide a first-person perspective of the wearer's activities involving simultaneous interactions with multiple objects. In this work we propo se the task of weakly-supervised Narration-based Video Object Segmentation (NVOS). Given an egocentric video clip and a narration of the wearer's activities our aim is to segment object instances mentioned in the narration without using any spatial annotations during training. Existing weakly-supervised video object gr ounding methods typically yield bounding boxes for referred objects. In contrast we propose ROSA a weakly-supervised pixel-level grounding framework learning al ignments between referred objects and segmentation mask proposals. Our model har nesses vision-language models pre-trained on image-text pairs to embed region ma sks and object phrases. During training we combine (a) a video-narration contras tive loss that implicitly supervises the alignment between regions and phrases a nd (b) a region-phrase contrastive loss based on inferred latent alignments. To address the lack of annotated NVOS datasets in egocentric videos we create a new

evaluation benchmark VISOR-NVOS leveraging existing annotations of segmentation masks from VISOR alongside 14.6k newly-collected object-based video clip narrations. Our approach achieves state-of-the-art zero-shot pixel-level grounding performance compared to strong baselines under similar supervision. Additionally we demonstrate generalization capabilities for zero-shot video object grounding on YouCook2 a third-person instructional video dataset.
****
Rewrite the Stars
Xu Ma, Xiyang Dai, Yue Bai, Yizhou Wang, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5694-5703
Recent studies have drawn attention to the untapped potential of the "star operation" (element-wise multiplication) in network design. While intuitive explanations abound the foundational rationale behind its application remains largely unexplored. Our study attempts to reveal the star operation's ability of mapping inputs into high-dimensional non-linear feature spaces--akin to kernel tricks--without widening the network. We further introduce StarNet a simple yet powerful prototype demonstrating impressive performance and low latency under compact network structure and efficient budget. Like stars in the sky the star operation appears unremarkable but holds a vast universe of potential. Our work encourages further exploration across tasks with codes available at https://github.com/ma-xu/Rewrite-the-Stars.
****
Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images
Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, Yanfeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11375-11385
Recent advancements in large-scale visual-language pre-trained models have led to significant progress in zero-/few-shot anomaly detection within natural image domains. However the substantial domain divergence between natural and medical images limits the effectiveness of these methodologies in medical anomaly detection. This paper introduces a novel lightweight multi-level adaptation and comparison framework to repurpose the CLIP model for medical anomaly detection. Our approach integrates multiple residual adapters into the pre-trained visual encoder enabling a stepwise enhancement of visual features across different levels. This multi-level adaptation is guided by multi-level pixel-wise visual-language feature alignment loss functions which recalibrate the model's focus from object semantics in natural imagery to anomaly identification in medical images. The adapted features exhibit improved generalization across various medical data types even in zero-shot scenarios where the model encounters unseen medical modalities and anatomical regions during training. Our experiments on medical anomaly detection benchmarks demonstrate that our method significantly surpasses current state-of-the-art models with an average AUC improvement of 6.24% and 7.33% for anomaly classification 2.03% and 2.37% for anomaly segmentation under the zero-shot and few-shot settings respectively. Source code is available at: https://github.com/MediaBrain-SJTU/MVFA-AD
****
AV-RIR: Audio-Visual Room Impulse Response Estimation
Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, Dinesh Manocha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27164-27175
Accurate estimation of Room Impulse Response (RIR) which captures an environment's acoustic properties is important for speech processing and AR/VR applications. We propose AV-RIR a novel multi-modal multi-task learning approach to accurately estimate the RIR from a given reverberant speech signal and the visual cues of its corresponding environment. AV-RIR builds on a novel neural codec-based architecture that effectively captures environment geometry and materials properties and solves speech dereverberation as an auxiliary task by using multi-task learning. We also propose Geo-Mat features that augment material information into visual cues and CRIP that improves late reverberation components in the estimated

RIR via image-to-RIR retrieval by 86%. Empirical results show that AV-RIR quantitatively outperforms previous audio-only and visual-only approaches by achieving 36% - 63% improvement across various acoustic metrics in RIR estimation. Additionally it also achieves higher preference scores in human evaluation. As an auxiliary benefit dereverbed speech from AV-RIR shows competitive performance with the state-of-the-art in various spoken language processing tasks and outperforms reverberation time error score in the real-world AVSpeech dataset. Qualitative examples of both synthesized reverberant speech and enhanced speech are available online https://www.youtube.com/watch?v=tTsKhviukAE.
****
Depth-aware Test-Time Training for Zero-shot Video Object Segmentation
Weihuang Liu, Xi Shen, Haolun Li, Xiuli Bi, Bo Liu, Chi-Man Pun, Xiaodong Cun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19218-19227
Zero-shot Video Object Segmentation (ZSVOS) aims at segmenting the primary moving object without any human annotations. Mainstream solutions mainly focus on learning a single model on large-scale video datasets which struggle to generalize to unseen videos. In this work we introduce a test-time training (TTT) strategy to address the problem. Our key insight is to enforce the model to predict consistent depth during the TTT process. In detail we first train a single network to perform both segmentation and depth prediction tasks. This can be effectively learned with our specifically designed depth modulation layer. Then for the TTT process the model is updated by predicting consistent depth maps for the same frame under different data augmentations. In addition we explore different TTT weight update strategies. Our empirical results suggest that the momentum-based weight initialization and looping-based training scheme lead to more stable improvements. Experiments show that the proposed method achieves clear improvements on ZSVOS. Our proposed video TTT strategy provides significant superiority over state-of-the-art TTT methods. Our code is available at: https://nifangbaage.github.io/DATTT/.
****
Dual-Consistency Model Inversion for Non-Exemplar Class Incremental Learning
Zihuan Qiu, Yi Xu, Fanman Meng, Hongliang Li, Linfeng Xu, Qingbo Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24025-24035
Non-exemplar class incremental learning (NECIL) aims to continuously assimilate new knowledge without forgetting previously acquired ones when historical data are unavailable. One of the generative NECIL methods is to invert the images of old classes for joint training. However these synthetic images suffer significant domain shifts compared with real data hampering the recognition of old classes. In this paper we present a novel method termed Dual-Consistency Model Inversion (DCMI) to generate better synthetic samples of old classes through two pivotal consistency alignments: (1) the semantic consistency between the synthetic images and the corresponding prototypes and (2) domain consistency between synthetic and real images of new classes. Besides we introduce Prototypical Routing (PR) to provide task-prior information and generate unbiased and accurate predictions. Our comprehensive experiments across diverse datasets consistently showcase the superiority of our method over previous state-of-the-art approaches.
****
RMem: Restricted Memory Banks Improve Video Object Segmentation
Junbao Zhou, Ziqi Pang, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18602-18611
With recent video object segmentation (VOS) benchmarks evolving to challenging scenarios we revisit a simple but overlooked strategy: restricting the size of memory banks. This diverges from the prevalent practice of expanding memory banks to accommodate extensive historical information. Our specially designed "memory deciphering" study offers a pivotal insight underpinning such a strategy: expanding memory banks while seemingly beneficial actually increases the difficulty for VOS modules to decode relevant features due to the confusion from redundant information. By restricting memory banks to a limited number of essential frames w

e achieve a notable improvement in VOS accuracy. This process balances the importance and freshness of frames to maintain an informative memory bank within a bounded capacity. Additionally restricted memory banks reduce the training-inference discrepancy in memory lengths compared with continuous expansion. This fosters new opportunities in temporal reasoning and enables us to introduce the previously overlooked "temporal positional embedding." Finally our insights are embodied in "RMem" ("R" for restricted) a simple yet effective VOS modification that excels at challenging VOS scenarios and establishes new state of the art for object state changes (VOST dataset) and long videos (the Long Videos dataset). Our codes are available at https://github.com/Restricted-Memory/RMemand our demo can be watched on https://youtu.be/V3tCFQsJrrM.
****
Not All Prompts Are Secure: A Switchable Backdoor Attack Against Pre-trained Vision Transfomers
Sheng Yang, Jiawang Bai, Kuofeng Gao, Yong Yang, Yiming Li, Shu-Tao Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24431-24441
Given the power of vision transformers a new learning paradigm pre-training and then prompting makes it more efficient and effective to address downstream visual recognition tasks. In this paper we identify a novel security threat towards such a paradigm from the perspective of backdoor attacks. Specifically an extra prompt token called the switch token in this work can turn the backdoor mode on i.e. converting a benign model into a backdoored one. Once under the backdoor mode a specific trigger can force the model to predict a target class. It poses a severe risk to the users of cloud API since the malicious behavior can not be activated and detected under the benign mode thus making the attack very stealthy. To attack a pre-trained model our proposed attack named SWARM learns a trigger and prompt tokens including a switch token. They are optimized with the clean loss which encourages the model always behaves normally even the trigger presents and the backdoor loss that ensures the backdoor can be activated by the trigger when the switch is on. Besides we utilize the cross-mode feature distillation to reduce the effect of the switch token on clean samples. The experiments on diverse visual recognition tasks confirm the success of our switchable backdoor attack i.e. achieving 95%+ attack success rate and also being hard to be detected and removed. Our code is available at https://github.com/20000yshust/SWARM.
****
PairDETR : Joint Detection and Association of Human Bodies and Faces
Ammar Ali, Georgii Gaikov, Denis Rybalchenko, Alexander Chigorin, Ivan Laptev, Sergey Zagoruyko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 423-432
Image and video analysis requires not only accurate object but also the understanding of relationships among detected objects. Common solutions to relation modeling typically resort to stand-alone object detectors followed by non-differentiable post-processing techniques. Recently introduced detection transformers (DETR) perform end-to-end object detection based on a bipartite matching loss. Such methods however lack the ability to jointly detect objects and resolve object associations. In this paper we build on the DETR approach and extend it to the joint detection of objects and their relationships by introducing an approximated bipartite matching. While our method can generalize to an arbitrary number of objects we here focus on the modeling of object pairs and their relations. In particular we apply our method PairDETR to the problem of detecting human bodies and faces and associating them for the same person. Our approach not only eliminates the need for hand-designed post-processing but also achieves excellent results for body-face associations. We evaluate PairDETR on the challenging CrowdHuman and CityPersons datasets and demonstrate a large improvement over the state of the art. Our training code and pre-trained models are available online.
****
PortraitBooth: A Versatile Portrait Model for Fast Identity-preserved Personalization
Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei L

in, Taisong Jin, Chengjie Wang, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27080-27090
Recent advancements in personalized image generation using diffusion models have been noteworthy. However existing methods suffer from inefficiencies due to the requirement for subject-specific fine-tuning. This computationally intensive process hinders efficient deployment limiting practical usability. Moreover these methods often grapple with identity distortion and limited expression diversity. In light of these challenges we propose PortraitBooth an innovative approach designed for high efficiency robust identity preservation and expression-editable text-to-image generation without the need for fine-tuning. PortraitBooth leverages subject embeddings from a face recognition model for personalized image generation without fine-tuning. It eliminates computational overhead and mitigates identity distortion. The introduced dynamic identity preservation strategy further ensures close resemblance to the original image identity. Moreover PortraitBooth incorporates emotion-aware cross-attention control for diverse facial expressions in generated images supporting text-driven expression editing. Its scalability enables efficient and high-quality image creation including multi-subject generation. Extensive results demonstrate superior performance over other state-of-the-art methods in both single and multiple image generation scenarios.
****
Learn from View Correlation: An Anchor Enhancement Strategy for Multi-view Clustering
Suyuan Liu, Ke Liang, Zhibin Dong, Siwei Wang, Xihong Yang, Sihang Zhou, En Zhu, Xinwang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26151-26161
In recent years anchor-based methods have achieved promising progress in multi-view clustering. The performances of these methods are significantly affected by the quality of the anchors. However the anchors generated by previous works solely rely on single-view information ignoring the correlation among different views. In particular we observe that similar patterns are more likely to exist between similar views so such correlation information can be leveraged to enhance the quality of the anchors which is also omitted. To this end we propose a novel plug-and-play anchor enhancement strategy through view correlation for multi-view clustering. Specifically we construct a view graph based on aligned initial anchor graphs to explore inter-view correlations. By learning from view correlation we enhance the anchors of the current view using the relationships between anchors and samples on neighboring views thereby narrowing the spatial distribution of anchors on similar views. Experimental results on seven datasets demonstrate the superiority of our proposed method over other existing methods. Furthermore extensive comparative experiments validate the effectiveness of the proposed anchor enhancement module when applied to various anchor-based methods.
****
SportsSloMo: A New Benchmark and Baselines for Human-centric Video Frame Interpolation
Jiaben Chen, Huaizu Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6475-6486
Human-centric video frame interpolation has great potential for enhancing entertainment experiences and finding commercial applications in sports analysis industry e.g. synthesizing slow-motion videos. Although there are multiple benchmark datasets available for video frame interpolation in the community none of them is dedicated to human-centric scenarios. To bridge this gap we introduce SportsSloMo a benchmark featuring over 130K high-resolution slow-motion sports video clips totaling over 1M video frames sourced from YouTube. We re-train several state-of-the-art methods on our benchmark and we observed a noticeable decrease in their accuracy compared to other datasets. This highlights the difficulty of our benchmark and suggests that it poses significant challenges even for the best-performing methods as human bodies are highly deformable and occlusions are frequent in sports videos. To tackle these challenges we propose human-aware loss terms where we add auxiliary supervision for human segmentation in panoptic settings and keypoints detection. These loss terms are model-agnostic and can be easily p

lugged into any video frame interpolation approach. Experimental results validat
e the effectiveness of our proposed human-aware loss terms leading to consistent
 performance improvement over existing models. The dataset and code can be found
 at: https://neu-vi.github.io/SportsSlomo/ https://neu-vi.github.io/SportsSlomo
/.
****
APSeg: Auto-Prompt Network for Cross-Domain Few-Shot Semantic Segmentation
Weizhao He, Yang Zhang, Wei Zhuo, Linlin Shen, Jiaqi Yang, Songhe Deng, Liang Su
n; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2024, pp. 23762-23772
Few-shot semantic segmentation (FSS) endeavors to segment unseen classes with on
ly a few labeled samples. Current FSS methods are commonly built on the assumpti
on that their training and application scenarios share similar domains and their
 performances degrade significantly while applied to a distinct domain. To this
end we propose to leverage the cutting-edge foundation model the Segment Anythin
g Model (SAM) for generalization enhancement. The SAM however performs unsatisfa
ctorily on domains that are distinct from its training data which primarily comp
rise natural scene images and it does not support automatic segmentation of spec
ific semantics due to its interactive prompting mechanism. In our work we introd
uce APSeg a novel auto-prompt network for cross-domain few-shot semantic segment
ation (CD-FSS) which is designed to be auto-prompted for guiding cross-domain se
gmentation. Specifically we propose a Dual Prototype Anchor Transformation (DPAT
) module that fuses pseudo query prototypes extracted based on cycle-consistency
 with support prototypes allowing features to be transformed into a more stable
domain-agnostic space. Additionally a Meta Prompt (MPG) module is introduced to
automatically generate prompt embeddings eliminating the need for manual visual
prompts. We build an efficient model which can be applied directly to target dom
ains without fine-tuning. Extensive experiments on four cross-domain datasets sh
ow that our model outperforms the state-of-the-art CD-FSS method by 5.24% and 3.
10% in average accuracy on 1-shot and 5-shot settings respectively.
****
Text2HOI: Text-guided 3D Motion Generation for Hand-Object Interaction
Junuk Cha, Jihyeon Kim, Jae Shin Yoon, Seungryul Baek; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1577-
1585
This paper introduces the first text-guided work for generating the sequence of
hand-object interaction in 3D. The main challenge arises from the lack of labele
d data where existing ground-truth datasets are nowhere near generalizable in in
teraction type and object category which inhibits the modeling of diverse 3D han
d-object interaction with the correct physical implication (e.g. contacts and se
mantics) from text prompts. To address this challenge we propose to decompose th
e interaction generation task into two subtasks: hand-object contact generation;
 and hand-object motion generation. For contact generation a VAE-based network t
akes as input a text and an object mesh and generates the probability of contact
s between the surfaces of hands and the object during the interaction. The netwo
rk learns a variety of local geometry structure of diverse objects that is indep
endent of the objects' category and thus it is applicable to general objects. Fo
r motion generation a Transformer-based diffusion model utilizes this 3D contact
 map as a strong prior for generating physically plausible hand-object motion as
 a function of text prompts by learning from the augmented labeled dataset; wher
e we annotate text labels from many existing 3D hand and object motion data. Fin
ally we further introduce a hand refiner module that minimizes the distance betw
een the object surface and hand joints to improve the temporal stability of the
object-hand contacts and to suppress the penetration artifacts. In the experimen
ts we demonstrate that our method can generate more realistic and diverse intera
ctions compared to other baseline methods. We also show that our method is appli
cable to unseen objects. We will release our model and newly labeled data as a s
trong foundation for future research. Codes and data are available in: https://g
ithub.com/JunukCha/Text2HOI.
****

Zero-TPrune: Zero-Shot Token Pruning through Leveraging of the Attention Graph i
n Pre-Trained Transformers
Hongjie Wang, Bhishma Dedhia, Niraj K. Jha; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16070-16079
Deployment of Transformer models on edge devices is becoming increasingly challe
nging due to the exponentially growing inference cost that scales quadratically
with the number of tokens in the input sequence. Token pruning is an emerging so
lution to address this challenge due to its ease of deployment on various Transf
ormer backbones. However most token pruning methods require computationally expe
nsive fine-tuning which is undesirable in many edge deployment cases. In this wo
rk we propose Zero-TPrune the first zero-shot method that considers both the imp
ortance and similarity of tokens in performing token pruning. It leverages the a
ttention graph of pre-trained Transformer models to produce an importance distri
bution for tokens via our proposed Weighted Page Rank (WPR) algorithm. This dist
ribution further guides token partitioning for efficient similarity-based prunin
g. Due to the elimination of the fine-tuning overhead Zero-TPrune can prune larg
e models at negligible computational cost switch between different pruning confi
gurations at no computational cost and perform hyperparameter tuning efficiently
. We evaluate the performance of Zero-TPrune on vision tasks by applying it to v
arious vision Transformer backbones and testing them on ImageNet. Without any fi
ne-tuning Zero-TPrune reduces the FLOPs cost of DeiT-S by 34.7% and improves its
 throughput by 45.3% with only 0.4% accuracy loss. Compared with state-of-the-ar
t pruning methods that require fine-tuning Zero-TPrune not only eliminates the n
eed for fine-tuning after pruning but also does so with only 0.1% accuracy loss.
 Compared with state-of-the-art fine-tuning-free pruning methods Zero-TPrune red
uces accuracy loss by up to 49% with the same or higher throughput.
****
Enhancing Visual Continual Learning with Language-Guided Supervision
Bolin Ni, Hongbo Zhao, Chenghao Zhang, Ke Hu, Gaofeng Meng, Zhaoxiang Zhang, Shi
ming Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2024, pp. 24068-24077
Continual learning (CL) aims to empower models to learn new tasks without forget
ting previously acquired knowledge. Most prior works concentrate on the techniqu
es of architectures replay data regularization etc. However the category name of
 each class is largely neglected. Existing methods commonly utilize the one-hot
labels and randomly initialize the classifier head. We argue that the scarce sem
antic information conveyed by the one-hot labels hampers the effective knowledge
 transfer across tasks. In this paper we revisit the role of the classifier head
 within the CL paradigm and replace the classifier with semantic knowledge from
pretrained language models (PLMs). Specifically we use PLMs to generate semantic
 targets for each class which are frozen and serve as supervision signals during
 training. Such targets fully consider the semantic correlation between all clas
ses across tasks. Empirical studies show that our approach mitigates forgetting
by alleviating representation drifting and facilitating knowledge transfer acros
s tasks. The proposed method is simple to implement and can seamlessly be plugge
d into existing methods with negligible adjustments. Extensive experiments based
 on eleven mainstream baselines demonstrate the effectiveness and generalizabili
ty of our approach to various protocols. For example under the class-incremental
 learning setting on ImageNet-100 our method significantly improves the Top-1 ac
curacy by 3.2% to 6.1% while reducing the forgetting rate by 2.6% to 13.1%.
****
MACE: Mass Concept Erasure in Diffusion Models
Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, Adams Wai-Kin Kong; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 6430-6440
The rapid expansion of large-scale text-to-image diffusion models has raised gro
wing concerns regarding their potential misuse in creating harmful or misleading
 content. In this paper we introduce MACE a finetuning framework for the task of
 mass concept erasure. This task aims to prevent models from generating images t
hat embody unwanted concepts when prompted. Existing concept erasure methods are

typically restricted to handling fewer than five concepts simultaneously and st ruggle to find a balance between erasing concept synonyms (generality) and maint aining unrelated concepts (specificity). In contrast MACE differs by successfull y scaling the erasure scope up to 100 concepts and by achieving an effective bal ance between generality and specificity. This is achieved by leveraging closed-f orm cross-attention refinement along with LoRA finetuning collectively eliminati ng the information of undesirable concepts. Furthermore MACE integrates multiple LoRAs without mutual interference. We conduct extensive evaluations of MACE aga inst prior methods across four different tasks: object erasure celebrity erasure explicit content erasure and artistic style erasure. Our results reveal that MA CE surpasses prior methods in all evaluated tasks. Code is available at https:// github.com/Shilin-LU/MACE.
****
DIBS: Enhancing Dense Video Captioning with Unlabeled Videos via Pseudo Boundary Enrichment and Online Refinement
Hao Wu, Huabin Liu, Yu Qiao, Xiao Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18699-18708
We present Dive Into the Boundaries (DIBS) a novel pretraining framework for den se video captioning (DVC) that elaborates on improving the quality of the genera ted event captions and their associated pseudo event boundaries from unlabeled v ideos. By leveraging the capabilities of diverse large language models (LLMs) we generate rich DVC-oriented caption candidates and optimize the corresponding ps eudo boundaries under several meticulously designed objectives considering diver sity event-centricity temporal ordering and coherence. Moreover we further intro duce a novel online boundary refinement strategy that iteratively improves the q uality of pseudo boundaries during training. Comprehensive experiments have been conducted to examine the effectiveness of the proposed technique components. By leveraging a substantial amount of unlabeled video data such as HowTo100M we ac hieve a remarkable advancement on standard DVC datasets like YouCook2 and Activi tyNet. We outperform the previous state-of-the-art Vid2Seq across a majority of metrics achieving this with just 0.4% of the unlabeled video data used for pre-t raining by Vid2Seq.
****
PeLK: Parameter-efficient Large Kernel ConvNets with Peripheral Convolution
Honghao Chen, Xiangxiang Chu, Yongjian Ren, Xin Zhao, Kaiqi Huang; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 5557-5567
Recently some large kernel convnets strike back with appealing performance and e fficiency. However given the square complexity of convolution scaling up kernels can bring about an enormous amount of parameters and the proliferated parameter s can induce severe optimization problem. Due to these issues current CNNs compr omise to scale up to 51x51 in the form of stripe convolution (i.e. 51x5+5x51) an d start to saturate as the kernel size continues growing. In this paper we delve into addressing these vital issues and explore whether we can continue scaling up kernels for more performance gains. Inspired by human vision we propose a hum an-like peripheral convolution that efficiently reduces over 90% parameter count of dense grid convolution through parameter sharing and manage to scale up kern el size to extremely large. Our peripheral convolution behaves highly similar to human reducing the complexity of convolution from O(K^2) to O(logK) without bac kfiring performance. Built on this we propose Parameter-efficient Large Kernel N etwork (PeLK). Our PeLK outperforms modern vision Transformers and ConvNet archi tectures like Swin ConvNeXt RepLKNet and SLaK on various vision tasks including ImageNet classification semantic segmentation on ADE20K and object detection on MS COCO. For the first time we successfully scale up the kernel size of CNNs to an unprecedented 101x101 and demonstrate consistent improvements.
****
AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation
Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, Zhongang Cai; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1834

-1843
Expressive human pose and shape estimation (a.k.a. 3D whole-body mesh recovery) involves the human body hand and expression estimation. Most existing methods have tackled this task in a two-stage manner first detecting the human body part with an off-the-shelf detection model and then inferring the different human body parts individually. Despite the impressive results achieved these methods suffer from 1) loss of valuable contextual information via cropping 2) introducing distractions and 3) lacking inter-association among different persons and body parts inevitably causing performance degradation especially for crowded scenes. To address these issues we introduce a novel all-in-one-stage framework AiOS for multiple expressive human pose and shape recovery without an additional human detection step. Specifically our method is built upon DETR which treats multi-person whole-body mesh recovery task as a progressive set prediction problem with various sequential detection. We devise the decoder tokens and extend them to our task. Specifically we first employ a human token to probe a human location in the image and encode global features for each instance which provides a coarse location for the later transformer block. Then we introduce a joint-related token to probe the human joint in the image and encoder a fine-grained local feature which collaborates with the global feature to regress the whole-body mesh. This straightforward but effective model outperforms previous state-of-the-art methods by a 9 reduction in NMVE on AGORA a 30 reduction in PVE on EHF a 10 reduction in PVE on ARCTIC and a 3 reduction in PVE on EgoBody.
****
SOK-Bench: A Situated Video Reasoning Benchmark with Aligned Open-World Knowledge
Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13384-13394
Reasoning from visual dynamics scenes has many real world applications. However existing video reasoning benchmarks are still inadequate since they were mainly designed for factual or situated reasoning and rarely involve broader knowledge in the real world. Our work aims to delve deeper into reasoning evaluations specifically within dynamic open-world and structured context knowledge. We propose a new benchmark (SOK-Bench) consisting of 44K questions and 10K situations with instance-level annotations depicted in the videos. The reasoning process is required to understand and apply situated knowledge and general knowledge for problem-solving. To create such a dataset we propose an automatic and scalable generation method to generate question-answer pairs knowledge graphs and rationales by instructing the combinations of LLMs and MLLMs. Concretely we first extract observable situated entities relations and processes from videos for situated knowledge and then extend to open-world knowledge beyond the visible content. The task generation is facilitated through multiple dialogues as iterations and subsequently corrected and refined by our designed self-promptings and demonstrations. With a corpus of both explicit situated facts and implicit commonsense we generate associated question-answer pairs and reasoning processes finally followed by manual reviews for quality assurance. We evaluated recent mainstream large vision language models on the benchmark and found several insightful conclusions. For more information please refer to our benchmark at www.bobbywu.com/SOKBench.
****
LORS: Low-rank Residual Structure for Parameter-Efficient Network Stacking
Jialin Li, Qiang Nie, Weifu Fu, Yuhuan Lin, Guangpin Tao, Yong Liu, Chengjie Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15866-15876
Deep learning models particularly those based on transformers often employ numerous stacked structures which possess identical architectures and perform similar functions. While effective this stacking paradigm leads to a substantial increase in the number of parameters pos- ing challenges for practical applications. In today's land- scape of increasingly large models stacking depth can even reach dozens further exacerbating this issue. To miti- gate this problem we introduce LORS (LOw-rank Residual Structure). LORS allows stacked modules to share the ma

jority of parameters requiring a much smaller num- ber of unique ones per module to match or even surpass the performance of using entirely distinct ones thereby significantly reducing parameter usage. We validate our method by applying it to the stacked decoders of a query- based object detector and conduct extensive experiments on the widely used MS COCO dataset. Experimental re- sults demonstrate the effectiveness of our method as even with a 70% reduction in the parameters of the decoder our method still enables the model to achieve comparable or even better performance than its original.
****
Design2Cloth: 3D Cloth Generation from 2D Masks
Jiali Zheng, Rolandos Alexandros Potamias, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1748-1758
In recent years there has been a significant shift in the field of digital avatar research towards modeling animating and reconstructing clothed human representations as a key step towards creating realistic avatars. However current 3D cloth generation methods are garment specific or trained completely on synthetic data hence lacking fine details and realism. In this work we make a step towards automatic realistic garment design and propose Design2Cloth a high fidelity 3D generative model trained on a real world dataset from more than 2000 subject scans. To provide vital contribution to the fashion industry we developed a user-friendly adversarial model capable of generating diverse and detailed clothes simply by drawing a 2D cloth mask. Under a series of both qualitative and quantitative experiments we showcase that Design2Cloth outperforms current state-of-the-art cloth generative models by a large margin. In addition to the generative properties of our network we showcase that the proposed method can be used to achieve high quality reconstructions from single in-the-wild images and 3D scans. Dataset code and pre-trained model will become publicly available.
****
Multi-modal In-Context Learning Makes an Ego-evolving Scene Text Recognizer
Zhen Zhao, Jingqun Tang, Chunhui Lin, Binghong Wu, Can Huang, Hao Liu, Xin Tan, Zhizhong Zhang, Yuan Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15567-15576
Scene text recognition (STR) in the wild frequently encounters challenges when coping with domain variations font diversity shape deformations etc. A straightforward solution is performing model fine-tuning tailored to a specific scenario but it is computationally intensive and requires multiple model copies for various scenarios. Recent studies indicate that large language models (LLMs) can learn from a few demonstration examples in a training-free manner termed "In-Context Learning" (ICL). Nevertheless applying LLMs as a text recognizer is unacceptably resource-consuming. Moreover our pilot experiments on LLMs show that ICL fails in STR mainly attributed to the insufficient incorporation of contextual information from diverse samples in the training stage. To this end we introduce E2STR a STR model trained with context-rich scene text sequences where the sequences are generated via our proposed in-context training strategy. E2STR demonstrates that a regular-sized model is sufficient to achieve effective ICL capabilities in STR. Extensive experiments show that E2STR exhibits remarkable training-free adaptation in various scenarios and outperforms even the fine-tuned state-of-the-art approaches on public benchmarks. The code is released at https://github.com/bytedance/E2STR.
****
Amodal Completion via Progressive Mixed Context Diffusion
Katherine Xu, Lingzhi Zhang, Jianbo Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9099-9109
Our brain can effortlessly recognize objects even when partially hidden from view. Seeing the visible of the hidden is called amodal completion; however this task remains a challenge for generative AI despite rapid progress. We propose to sidestep many of the difficulties of existing approaches which typically involve a two-step process of predicting amodal masks and then generating pixels. Our method involves thinking outside the box literally! We go outside the object bound

ing box to use its context to guide a pre-trained diffusion inpainting model and then progressively grow the occluded object and trim the extra background. We overcome two technical challenges: 1) how to be free of unwanted co-occurrence bias which tends to regenerate similar occluders and 2) how to judge if an amodal completion has succeeded. Our amodal completion method exhibits improved photorealistic completion results compared to existing approaches in numerous successful completion cases. And the best part? It doesn't require any special training or fine-tuning of models. Project page and code: https://k8xu.github.io/amodal/
****
Training Diffusion Models Towards Diverse Image Generation with Reinforcement Learning
Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, Zicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10844-10853
Diffusion models have demonstrated unprecedented capabilities in image generation. Yet they incorporate and amplify the data bias (e.g. gender age) from the original training set limiting the diversity of generated images. In this paper we propose a diversity-oriented fine-tuning method using reinforcement learning (RL) for diffusion models under the guidance of an image-set-based reward function. Specifically the proposed reward function denoted as Diversity Reward utilizes a set of generated images to evaluate the coverage of the current generative distribution w.r.t. the reference distribution represented by a set of unbiased images. Built on top of the probabilistic method of distribution discrepancy estimation Diversity Reward can measure the relative distribution gap with a small set of images efficiently. We further formulate the diffusion process as a multi-step decision-making problem (MDP) and apply policy gradient methods to fine-tune diffusion models by maximizing the Diversity Reward. The proposed rewards are validated on a post-sampling selection task where a subset of the most diverse images are selected based on Diversity Reward values. We also show the effectiveness of our RL fine-tuning framework on enhancing the diversity of image generation with different types of diffusion models including class-conditional models and text-conditional models e.g. StableDiffusion.
****
Diffusion 3D Features (Diff3F): Decorating Untextured Shapes with Distilled Semantic Features
Niladri Shekhar Dutt, Sanjeev Muralikrishnan, Niloy J. Mitra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4494-4504
We present Diff3F as a simple robust and class-agnostic feature descriptor that can be computed for untextured input shapes (meshes or point clouds). Our method distills diffusion features from image foundational models onto input shapes. Specifically we use the input shapes to produce depth and normal maps as guidance for conditional image synthesis. In the process we produce (diffusion) features in 2D that we subsequently lift and aggregate on the original surface. Our key observation is that even if the conditional image generations obtained from multi-view rendering of the input shapes are inconsistent the associated image features are robust and hence can be directly aggregated across views. This produces semantic features on the input shapes without requiring additional data or training. We perform extensive experiments on multiple benchmarks (SHREC'19 SHREC'20 FAUST and TOSCA) and demonstrate that our features being semantic instead of geometric produce reliable correspondence across both isometric and non-isometrically related shape families. Code is available via the project webpage at https://diff3f.github.io/
****
LASIL: Learner-Aware Supervised Imitation Learning For Long-term Microscopic Traffic Simulation
Ke Guo, Zhenwei Miao, Wei Jing, Weiwei Liu, Weizi Li, Dayang Hao, Jia Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15386-15395
Microscopic traffic simulation plays a crucial role in transportation engineerin

g by providing insights into individual vehicle behavior and overall traffic flo
w. However creating a realistic simulator that accurately replicates human drivi
ng behaviors in various traffic conditions presents significant challenges. Trad
itional simulators relying on heuristic models often fail to deliver accurate si
mulations due to the complexity of real-world traffic environments. Due to the c
ovariate shift issue existing imitation learning-based simulators often fail to
generate stable long-term simulations. In this paper we propose a novel approach
 called learner-aware supervised imitation learning to address the covariate shi
ft problem in multi-agent imitation learning. By leveraging a variational autoen
coder simultaneously modeling the expert and learner state distribution our appr
oach augments expert states such that the augmented state is aware of learner st
ate distribution. Our method applied to urban traffic simulation demonstrates si
gnificant improvements over existing state-of-the-art baselines in both short-te
rm microscopic and long-term macroscopic realism when evaluated on the real-worl
d dataset pNEUMA.
****
Revamping Federated Learning Security from a Defender's Perspective: A Unified D
efense with Homomorphic Encrypted Data Space
K Naveen Kumar, Reshmi Mitra, C Krishna Mohan; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24387-24397
Federated Learning (FL) facilitates clients to collaborate on training a shared
machine learning model without exposing individual private data. Nonetheless FL
remains susceptible to utility and privacy attacks notably evasion data poisonin
g and model inversion attacks compromising the system's efficiency and data priv
acy. Existing FL defenses are often specialized to a particular single attack la
cking generality and a comprehensive defender's perspective. To address these ch
allenges we introduce Federated Cryptography Defense (FCD) a unified single fram
ework aligning with the defender's perspective. FCD employs row-wise transpositi
on cipher based data encryption with a secret key to counter both evasion black-
box data poisoning and model inversion attacks. The crux of FCD lies in transfer
ring the entire learning process into an encrypted data space and using a novel
distillation loss guided by the Kullback-Leibler (KL) divergence. This measure c
ompares the probability distributions of the local pretrained teacher model's pr
edictions on normal data and the local student model's predictions on the same d
ata in FCD's encrypted form. By working within this encrypted space FCD eliminat
es the need for decryption at the server resulting in reduced computational comp
lexity. We demonstrate the practical feasibility of FCD and apply it to defend a
gainst evasion utility attack on benchmark datasets (GTSRB KBTS CIFAR10 and EMNI
ST). We further extend FCD for defending against model inversion attack in split
 FL on the CIFAR100 dataset. Our experiments across the diverse attack and FL se
ttings demonstrate practical feasibility and robustness against utility evasion
(impact >30) and privacy attacks (MSE >73) compared to the second best method.
****
A Dynamic Kernel Prior Model for Unsupervised Blind Image Super-Resolution
Zhixiong Yang, Jingyuan Xia, Shengxi Li, Xinghua Huang, Shuanghui Zhang, Zhen Li
u, Yaowen Fu, Yongxiang Liu; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2024, pp. 26046-26056
Deep learning-based methods have achieved significant successes on solving the b
lind super-resolution (BSR) problem. However most of them request supervised pre
-training on labelled datasets. This paper proposes an unsupervised kernel estim
ation model named dynamic kernel prior (DKP) to realize an unsupervised and pre-
training-free learning-based algorithm for solving the BSR problem. DKP can adap
tively learn dynamic kernel priors to realize real-time kernel estimation and th
ereby enables superior HR image restoration performances. This is achieved by a
Markov chain Monte Carlo sampling process on random kernel distributions. The le
arned kernel prior is then assigned to optimize a blur kernel estimation network
 which entails a network-based Langevin dynamic optimization strategy. These two
 techniques ensure the accuracy of the kernel estimation. DKP can be easily used
 to replace the kernel estimation models in the existing methods such as Double-
DIP and FKP-DIP or be added to the off-the-shelf image restoration model such as

diffusion model. In this paper we incorporate our DKP model with DIP and diffusion model referring to DIP-DKP and Diff-DKP for validations. Extensive simulations on Gaussian and motion kernel scenarios demonstrate that the proposed DKP model can significantly improve the kernel estimation with comparable runtime and memory usage leading to state-of-the-art BSR results. The code is available at https://github.com/XYLGroup/DKP.
****
Cinematic Behavior Transfer via NeRF-based Differentiable Filming
Xuekun Jiang, Anyi Rao, Jingbo Wang, Dahua Lin, Bo Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6723-6732
In the evolving landscape of digital media and video production the precise manipulation and reproduction of visual elements like camera movements and character actions are highly desired. Existing SLAM methods face limitations in dynamic scenes and human pose estimation often focuses on 2D projections neglecting 3D statuses. To address these issues we first introduce a reverse filming behavior estimation technique. It optimizes camera trajectories by leveraging NeRF as a differentiable renderer and refining SMPL tracks. We then introduce a cinematic transfer pipeline that is able to transfer various shot types to a new 2D video or a 3D virtual environment. The incorporation of 3D engine workflow enables superior rendering and control abilities which also achieves a higher rating in the user study.
****
SeaBird: Segmentation in Bird's View with Dice Loss Improves Monocular 3D Detection of Large Objects
Abhinav Kumar, Yuliang Guo, Xinyu Huang, Liu Ren, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10269-10280
Monocular 3D detectors achieve remarkable performance on cars and smaller objects. However their performance drops on larger objects leading to fatal accidents. Some attribute the failures to training data scarcity or the receptive field requirements of large objects. In this paper we highlight this understudied problem of generalization to large objects. We find that modern frontal detectors struggle to generalize to large objects even on nearly balanced datasets. We argue that the cause of failure is the sensitivity of depth regression losses to noise of larger objects. To bridge this gap we comprehensively investigate regression and dice losses examining their robustness under varying error levels and object sizes. We mathematically prove that the dice loss leads to superior noise-robustness and model convergence for large objects compared to regression losses for a simplified case. Leveraging our theoretical insights we propose SeaBird (Segmentation in Bird's View) as the first step towards generalizing to large objects. SeaBird effectively integrates BEV segmentation on foreground objects for 3D detection with the segmentation head trained with the dice loss. SeaBird achieves SoTA results on the KITTI-360 leaderboard and improves existing detectors on the nuScenes leaderboard particularly for large objects.
****
Text-Driven Image Editing via Learnable Regions
Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7059-7068
Language has emerged as a natural interface for image editing. In this paper we introduce a method for region-based image editing driven by textual prompts without the need for user-provided masks or sketches. Specifically our approach leverages an existing pre-trained text-to-image model and introduces a bounding box generator to identify the editing regions that are aligned with the textual prompts. We show that this simple approach enables flexible editing that is compatible with current image generation models and is able to handle complex prompts featuring multiple objects complex sentences or lengthy paragraphs. We conduct an extensive user study to compare our method against state-of-the-art methods. The experiments demonstrate the competitive performance of our method in manipulati

ng images with high fidelity and realism that correspond to the provided languag
e descriptions. Our project webpage can be found at: https://yuanzelin.me/Learna
bleRegions_page.
****
Relation Rectification in Diffusion Model
Yinwei Wu, Xingyi Yang, Xinchao Wang; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7685-7694
Despite their exceptional generative abilities large T2I diffusion models much l
ike skilled but careless artists often struggle with accurately depicting visual
 relationships between objects. This issue as we uncover through careful analysi
s arises from a misaligned text encoder that struggles to interpret specific rel
ationships and differentiate the logical order of associated objects. To resolve
 this we introduce a novel task termed Relation Rectification aiming to refine t
he model to accurately represent a given relationship it initially fails to gene
rate. To address this we propose an innovative solution utilizing a Heterogeneou
s Graph Convolutional Network (HGCN). It models the directional relationships be
tween relation terms and corresponding objects within the input prompts. Specifi
cally we optimize the HGCN on a pair of prompts with identical relational words
but reversed object orders supplemented by a few reference images. The lightweig
ht HGCN adjusts the text embeddings generated by the text encoder ensuring accur
ate reflection of the textual relation in the embedding space. Crucially our met
hod retains the parameters of the text encoder and diffusion model preserving th
e model's robust performance on unrelated descriptions. We validated our approac
h on a newly curated dataset of diverse relational data demonstrating both quant
itative and qualitative enhancements in generating images with precise visual re
lations. Project page: https://wuyinwei-hah.github.io/rrnet.github.io/ .
****
NOPE: Novel Object Pose Estimation from a Single Image
Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Yinlin Hu, Renaud Marlet
, Mathieu Salzmann, Vincent Lepetit; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2024, pp. 17923-17932
The practicality of 3D object pose estimation remains limited for many applicati
ons due to the need for prior knowledge of a 3D model and a training period for
new objects. To address this limitation we propose an approach that takes a sing
le image of a new object as input and predicts the relative pose of this object
in new images without prior knowledge of the object's 3D model and without requi
ring training time for new objects and categories. We achieve this by training a
 model to directly predict discriminative embeddings for viewpoints surrounding
the object. This prediction is done using a simple U-Net architecture with atten
tion and conditioned on the desired pose which yields extremely fast inference.
We compare our approach to state-of-the-art methods and show it outperforms them
 both in terms of accuracy and robustness.
****
Mocap Everyone Everywhere: Lightweight Motion Capture With Smartwatches and a He
ad-Mounted Camera
Jiye Lee, Hanbyul Joo; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2024, pp. 1091-1100
We present a lightweight and affordable motion capture method based on two smart
watches and a head-mounted camera. In contrast to the existing approaches that u
se six or more expert-level IMU devices our approach is much more cost-effective
 and convenient. Our method can make wearable motion capture accessible to every
one everywhere enabling 3D full-body motion capture in diverse environments. As
a key idea to overcome the extreme sparsity and ambiguities of sensor inputs wit
h different modalities we integrate 6D head poses obtained from the head-mounted
 cameras for motion estimation. To enable capture in expansive indoor and outdoo
r scenes we propose an algorithm to track and update floor level changes to defi
ne head poses coupled with a multi-stage Transformer-based regression module. We
 also introduce novel strategies leveraging visual cues of egocentric images to
further enhance the motion capture quality while reducing ambiguities. We demons
trate the performance of our method on various challenging scenarios including c

omplex outdoor environments and everyday motions including object interactions a
nd social interactions among multiple individuals.
****

Fast ODE-based Sampling for Diffusion Models in Around 5 Steps
Zhenyu Zhou, Defang Chen, Can Wang, Chun Chen; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7777-7786
Sampling from diffusion models can be treated as solving the corresponding ordin
ary differential equations (ODEs) with the aim of obtaining an accurate solution
 with as few number of function evaluations (NFE) as possible. Recently various
fast samplers utilizing higher-order ODE solvers have emerged and achieved bette
r performance than the initial first-order one. However these numerical methods
inherently result in certain approximation errors which significantly degrades s
ample quality with extremely small NFE (e.g. around 5). In contrast based on the
 geometric observation that each sampling trajectory almost lies in a two-dimens
ional subspace embedded in the ambient space we propose Approximate MEan-Directi
on Solver (AMED-Solver) that eliminates truncation errors by directly learning t
he mean direction for fast diffusion sampling. Besides our method can be easily
used as a plugin to further improve existing ODE-based samplers. Extensive exper
iments on image synthesis with the resolution ranging from 32 to 512 demonstrate
 the effectiveness of our method. With only 5 NFE we achieve 6.61 FID on CIFAR-1
0 10.74 FID on ImageNet 64x64 and 13.20 FID on LSUN Bedroom. Our code is availab
le at https://github.com/zju-pi/diff-sampler.
****

Dual-View Visual Contextualization for Web Navigation
Jihyung Kil, Chan Hee Song, Boyuan Zheng, Xiang Deng, Yu Su, Wei-Lun Chao; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 14445-14454
Automatic web navigation aims to build a web agent that can follow language inst
ructions to execute complex and diverse tasks on real-world websites. Existing w
ork primarily takes HTML documents as input which define the contents and action
 spaces (i.e. actionable elements and operations) of webpages. Nevertheless HTML
 documents may not provide a clear task-related context for each element making
it hard to select the right (sequence of) actions. In this paper we propose to c
ontextualize HTML elements through their "dual views" in webpage screenshots: ea
ch HTML element has its corresponding bounding box and visual content in the scr
eenshot. We build upon the insight---web developers tend to arrange task-related
 elements nearby on webpages to enhance user experiences---and propose to contex
tualize each element with its neighbor elements using both textual and visual fe
atures. The resulting representations of HTML elements are more informative for
the agent to take action. We validate our method on the recently released Mind2W
eb dataset which features diverse navigation domains and tasks on real-world web
sites. Our method consistently outperforms the baseline in all the scenarios inc
luding cross-task cross-website and cross-domain ones.
****

Language-driven Grasp Detection
An Dinh Vuong, Minh Nhat Vu, Baoru Huang, Nghia Nguyen, Hieu Le, Thieu Vo, Anh N
guyen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2024, pp. 17902-17912
Grasp detection is a persistent and intricate challenge with various industrial
applications. Recently many methods and datasets have been proposed to tackle th
e grasp detection problem. However most of them do not consider using natural la
nguage as a condition to detect the grasp poses. In this paper we introduce Gras
p-Anything++ a new language-driven grasp detection dataset featuring 1M samples
over 3M objects and upwards of 10M grasping instructions. We utilize foundation
models to create a large-scale scene corpus with corresponding images and grasp
prompts. We approach the language-driven grasp detection task as a conditional g
eneration problem. Drawing on the success of diffusion models in generative task
s and given that language plays a vital role in this task we propose a new langu
age-driven grasp detection method based on diffusion models. Our key contributio
n is the contrastive training objective which explicitly contributes to the deno

ising process to detect the grasp pose given the language instructions. We illus trate that our approach is theoretically supportive. The intensive experiments s how that our method outperforms state-of-the-art approaches and allows real-worl d robotic grasping. Finally we demonstrate our large-scale dataset enables zero-short grasp detection and is a challenging benchmark for future work.
****

Towards Modern Image Manipulation Localization: A Large-Scale Dataset and Novel Methods
Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, Lianwen Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2024, pp. 10781-10790
In recent years image manipulation localization has attracted increasing attenti on due to its pivotal role in ensuring social media security. However effectivel y identifying forged regions remains an open challenge. The high acquisition cos t and the severe scarcity of high-quality data are major factors hindering the p erformance improvement of modern image manipulation localization systems. To add ress this issue we propose a novel paradigm termed as CAAA to automatically and accurately annotate the manually forged images from the web at the pixel-level. We further propose a novel metric termed as QES to assist in filtering out unrel iable annotations. With CAAA and QES we construct a large-scale diverse and high -quality dataset comprising 123150 manually forged images with mask annotations. Furthermore we develop a new model termed as APSC-Net for accurate image manipu lation localization. According to extensive experiments our methods outperforms previous state-of-the-art methods our dataset significantly improves the perform ance of various models on the widely-used benchmarks. The dataset and codes are publicly available at https://github.com/qcf-568/MIML.
****

Mitigating Noisy Correspondence by Geometrical Structure Consistency Learning
Zihua Zhao, Mengxi Chen, Tianjie Dai, Jiangchao Yao, Bo Han, Ya Zhang, Yanfeng W ang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2024, pp. 27381-27390
Noisy correspondence that refers to mismatches in cross-modal data pairs is prev alent on human-annotated or web-crawled datasets. Prior approaches to leverage s uch data mainly consider the application of uni-modal noisy label learning witho ut amending the impact on both cross-modal and intra-modal geometrical structure s in multimodal learning. Actually we find that both structures are effective to discriminate noisy correspondence through structural differences when being wel l-established. Inspired by this observation we introduce a Geometrical Structure Consistency (GSC) method to infer the true correspondence. Specifically GSC ens ures the preservation of geometrical structures within and between modalities al lowing for the accurate discrimination of noisy samples based on structural diff erences. Utilizing these inferred true correspondence labels GSC refines the lea rning of geometrical structures by filtering out the noisy samples. Experiments across four cross-modal datasets confirm that GSC effectively identifies noisy s amples and significantly outperforms the current leading methods. Source code is available at https://github.com/MediaBrain-SJTU/GSC.
****

CLiC: Concept Learning in Context
Mehdi Safaee, Aryan Mikaeili, Or Patashnik, Daniel Cohen-Or, Ali Mahdavi-Amiri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2024, pp. 6924-6933
This paper addresses the challenge of learning a local visual pattern of an obje ct from one image and generating images depicting objects with that pattern. Lea rning a localized concept and placing it on an object in a target image is a non trivial task as the objects may have different orientations and shapes. Our appr oach builds upon recent advancements in visual concept learning. It involves acq uiring a visual concept (e.g. an ornament) from a source image and subsequently applying it to an object (e.g. a chair) in a target image. Our key idea is to pe rform in-context concept learning acquiring the local visual concept within the broader context of the objects they belong to. To localize the concept learning

we employ soft masks that contain both the concept within the mask and the surrounding image area. We demonstrate our approach through object generation within an image showcasing plausible embedding of in-context learned concepts. We also introduce methods for directing acquired concepts to specific locations within target images employing cross-attention mechanisms and establishing correspondences between source and target objects. The effectiveness of our method is demonstrated through quantitative and qualitative experiments along with comparisons against baseline techniques.
****
CAD-SIGNet: CAD Language Inference from Point Clouds using Layer-wise Sketch Instance Guided Attention
Mohammad Sadil Khan, Elona Dupont, Sk Aziz Ali, Kseniya Cherenkova, Anis Kacem, Djamila Aouada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4713-4722
Reverse engineering in the realm of Computer-Aided Design (CAD) has been a longstanding aspiration though not yet entirely realized. Its primary aim is to uncover the CAD process behind a physical object given its 3D scan. We propose CAD-SIGNet an end-to-end trainable and auto-regressive architecture to recover the design history of a CAD model represented as a sequence of sketch- and-extrusion from an input point cloud. Our model learns CAD visual-language representations by layer-wise cross-attention between point cloud and CAD language embedding. In particular a new Sketch instance Guided Attention (SGA) module is proposed in order to reconstruct the fine- grained details of the sketches. Thanks to its auto-regressive nature CAD-SIGNet not only reconstructs a unique full design history of the corresponding CAD model given an in- put point cloud but also provides multiple plausible design choices. This allows for an interactive reverse engineering scenario by providing designers with multiple next step choices along with the design process. Extensive experiments on publicly available CAD datasets showcase the effectiveness of our approach against existing baseline models in two settings namely full design history recovery and conditional auto-completion from point clouds.
****
Object Recognition as Next Token Prediction
Kaiyu Yue, Bor-Chun Chen, Jonas Geiping, Hengduo Li, Tom Goldstein, Ser-Nam Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16645-16656
We present an approach to pose object recognition as next token prediction. The idea is to apply a language decoder that auto-regressively predicts the text tokens from image embeddings to form labels. To ground this prediction process in auto-regression we customize a non-causal attention mask for the decoder incorporating two key features: modeling tokens from different labels to be independent and treating image tokens as a prefix. This masking mechanism inspires an efficient method -- one-shot sampling -- to simultaneously sample tokens of multiple labels in parallel and rank generated labels by their probabilities during inference. To further enhance the efficiency we propose a simple strategy to construct a compact decoder by simply discarding the intermediate blocks of a pretrained language model. This approach yields a decoder that matches the full model's performance while being notably more efficient. The code is available at https://github.com/kaiyuyue/nxtp.
****
CLIB-FIQA: Face Image Quality Assessment with Confidence Calibration
Fu-Zhao Ou, Chongyi Li, Shiqi Wang, Sam Kwong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1694-1704
Face Image Quality Assessment (FIQA) is pivotal for guaranteeing the accuracy of face recognition in unconstrained environments. Recent progress in deep quality-fitting-based methods that train models to align with quality anchors has shown promise in FIQA. However these methods heavily depend on a recognition model to yield quality anchors and indiscriminately treat the confidence of inaccurate anchors as equivalent to that of accurate ones during the FIQA model training leading to a fitting bottleneck issue. This paper seeks a solution by putting forwa

rd the Confidence-Calibrated Face Image Quality Assessment (CLIB-FIQA) approach underpinned by the synergistic interplay between the quality anchors and objective quality factors such as blur pose expression occlusion and illumination. Specifically we devise a joint learning framework built upon the vision-language alignment model which leverages the joint distribution with multiple quality factors to facilitate the quality fitting of the FIQA model. Furthermore to alleviate the issue of the model placing excessive trust in inaccurate quality anchors we propose a confidence calibration method to correct the quality distribution by exploiting to the fullest extent of these objective quality factors characterized as the merged-factor distribution during training. Experimental results on eight datasets reveal the superior performance of the proposed method.
****
DVMNet: Computing Relative Pose for Unseen Objects Beyond Hypotheses
Chen Zhao, Tong Zhang, Zheng Dang, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20485-20495
Determining the relative pose of an object between two images is pivotal to the success of generalizable object pose estimation. Existing approaches typically approximate the continuous pose representation with a large number of discrete pose hypotheses which incurs a computationally expensive process of scoring each hypothesis at test time. By contrast we present a Deep Voxel Matching Network (DVMNet) that eliminates the need for pose hypotheses and computes the relative object pose in a single pass. To this end we map the two input RGB images reference and query to their respective voxelized 3D representations. We then pass the resulting voxels through a pose estimation module where the voxels are aligned and the pose is computed in an end-to-end fashion by solving a least-squares problem. To enhance robustness we introduce a weighted closest voxel algorithm capable of mitigating the impact of noisy voxels. We conduct extensive experiments on the CO3D LINEMOD and Objaverse datasets demonstrating that our method delivers more accurate relative pose estimates for novel objects at a lower computational cost compared to state-of-the-art methods. Our code is released at: https://github.com/sailor-z/DVMNet.
****
Transcriptomics-guided Slide Representation Learning in Computational Pathology
Guillaume Jaume, Lukas Oldenburg, Anurag Vaidya, Richard J. Chen, Drew F.K. Williamson, Thomas Peeters, Andrew H. Song, Faisal Mahmood; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9632-9644
Self-supervised learning (SSL) has been successful in building patch embeddings of small histology images (e.g. 224 x 224 pixels) but scaling these models to learn slide embeddings from the entirety of giga-pixel whole-slide images (WSIs) remains challenging. Here we leverage complementary information from gene expression profiles to guide slide representation learning using multi-modal pre-training. Expression profiles constitute highly detailed molecular descriptions of a tissue that we hypothesize offer a strong task-agnostic training signal for learning slide embeddings. Our slide and expression (S+E) pretraining strategy called TANGLE employs modality-specific encoders the outputs of which are aligned via contrastive learning. TANGLE was pre-trained on samples from three different organs: liver (n=6597 S+E pairs) breast (n=1020) and lung (n=1012) from two different species (Homo sapiens and Rattus norvegicus). Across three independent test datasets consisting of 1265 breast WSIs 1946 lung WSIs and 4584 liver WSIs TANGLE shows significantly better few-shot performance compared to supervised and SSL baselines. When assessed using prototype-based classification and slide retrieval TANGLE also shows a substantial performance improvement over all baselines. Code available at https://github.com/mahmoodlab/TANGLE.
****
Predicated Diffusion: Predicate Logic-Based Attention Guidance for Text-to-Image Diffusion Models
Kota Sueyoshi, Takashi Matsubara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8651-8660

Diffusion models have achieved remarkable success in generating high-quality diverse and creative images. However in text-based image generation they often struggle to accurately capture the intended meaning of the text. For instance a specified object might not be generated or an adjective might incorrectly alter unintended objects. Moreover we found that relationships indicating possession between objects are frequently overlooked. Despite the diversity of users' intentions in text existing methods often focus on only some aspects of these intentions. In this paper we propose Predicated Diffusion a unified framework designed to more effectively express users' intentions. It represents the intended meaning as propositions using predicate logic and treats the pixels in attention maps as fuzzy predicates. This approach provides a differentiable loss function that offers guidance for the image generation process to better fulfill the propositions. Comparative evaluations with existing methods demonstrated that Predicated Diffusion excels in generating images faithful to various text prompts while maintaining high image quality as validated by human evaluators and pretrained image-text models.
****
MuRF: Multi-Baseline Radiance Fields
Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20041-20050
We present Multi-Baseline Radiance Fields (MuRF) a general feed-forward approach to solving sparse view synthesis under multiple different baseline settings (small and large baselines and different number of input views). To render a target novel view we discretize the 3D space into planes parallel to the target image plane and accordingly construct a target view frustum volume. Such a target volume representation is spatially aligned with the target view which effectively aggregates relevant information from the input views for high-quality rendering. It also facilitates subsequent radiance field regression with a convolutional network thanks to its axis-aligned nature. The 3D context modeled by the convolutional network enables our method to synthesis sharper scene structures than prior works. Our MuRF achieves state-of-the-art performance across multiple different baseline settings and diverse scenarios ranging from simple objects (DTU) to complex indoor and outdoor scenes (RealEstate10K and LLFF). We also show promising zero-shot generalization abilities on the Mip-NeRF 360 dataset demonstrating the general applicability of MuRF.
****
CLIP-BEVFormer: Enhancing Multi-View Image-Based BEV Detector with Ground Truth Flow
Chenbin Pan, Burhaneddin Yaman, Senem Velipasalar, Liu Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15216-15225
Autonomous driving stands as a pivotal domain in computer vision shaping the future of transportation. Within this paradigm the backbone of the system plays a crucial role in interpreting the complex environment. However a notable challenge has been the loss of clear supervision when it comes to Bird's Eye View elements. To address this limitation we introduce CLIP-BEVFormer a novel approach that leverages the power of contrastive learning techniques to enhance the multi-view image-derived BEV backbones with ground truth information flow. We conduct extensive experiments on the challenging nuScenes dataset and showcase significant and consistent improvements over the SOTA. Specifically CLIP-BEVFormer achieves an impressive 8.5% and 9.2% enhancement in terms of NDS and mAP respectively over the previous best BEV model on the 3D object detection task.
****
CLOVA: A Closed-LOop Visual Assistant with Tool Usage and Update
Zhi Gao, Yuntao Du, Xintong Zhang, Xiaojian Ma, Wenjuan Han, Song-Chun Zhu, Qing Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13258-13268
Utilizing large language models (LLMs) to compose off-the-shelf visual tools represents a promising avenue of research for developing robust visual assistants c

apable of addressing diverse visual tasks. However these methods often overlook the potential for continual learning typically by freezing the utilized tools th us limiting their adaptation to environments requiring new knowledge. To tackle this challenge we propose CLOVA a Closed-Loop Visual Assistant which operates wi thin a framework encompassing inference reflection and learning phases. During t he inference phase LLMs generate programs and execute corresponding tools to com plete assigned tasks. In the reflection phase a multimodal global-local reflecti on scheme analyzes human feedback to determine which tools require updating. Las tly the learning phase employs three flexible approaches to automatically gather training data and introduces a novel prompt tuning scheme to update the tools a llowing CLOVA to efficiently acquire new knowledge. Experimental findings demons trate that CLOVA surpasses existing tool-usage methods by 5% in visual question answering and multiple-image reasoning by 10% in knowledge tagging and by 20% in image editing. These results underscore the significance of the continual learn ing capability in general visual assistants.
****
Depth Prompting for Sensor-Agnostic Depth Estimation
Jin-Hwi Park, Chanhwi Jeong, Junoh Lee, Hae-Gon Jeon; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9859-9 869
Dense depth maps have been used as a key element of visual perception tasks. The re have been tremendous efforts to enhance the depth quality ranging from optimi zation-based to learning-based methods. Despite the remarkable progress for a lo ng time their applicability in the real world is limited due to systematic measu rement biases such as density sensing pattern and scan range. It is well-known t hat the biases make it difficult for these methods to achieve their generalizati on. We observe that learning a joint representation for input modalities (e.g. i mages and depth) which most recent methods adopt is sensitive to the biases. In this work we disentangle those modalities to mitigate the biases with prompt eng ineering. For this we design a novel depth prompt module to allow the desirable feature representation according to new depth distributions from either sensor t ypes or scene configurations. Our depth prompt can be embedded into foundation m odels for monocular depth estimation. Through this embedding process our method helps the pretrained model to be free from restraint of depth scan range and to provide absolute scale depth maps. We demonstrate the effectiveness of our metho d through extensive evaluations. Source code is publicly available at https://gi thub.com/JinhwiPark/DepthPrompting.
****
G3DR: Generative 3D Reconstruction in ImageNet
Pradyumna Reddy, Ismail Elezi, Jiankang Deng; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9655-9665
We introduce a novel 3D generative method Generative 3D Reconstruction (G3DR) in ImageNet capable of generating diverse and high-quality 3D objects from single images addressing the limitations of existing methods. At the heart of our frame work is a novel depth regularization technique that enables the generation of sc enes with high-geometric fidelity. G3DR also leverages a pretrained language-vis ion model such as CLIP to enable reconstruction in novel views and improve the v isual realism of generations. Additionally G3DR designs a simple but effective s ampling procedure to further improve the quality of generations. G3DR offers div erse and efficient 3D asset generation based on class or text conditioning. Desp ite its simplicity G3DR is able to beat state-of-theart methods improving over t hem by up to 22% in perceptual metrics and 90% in geometry scores while needing only half of the training time. Code is available at https://github.com/preddy5/ G3DR
****
MoML: Online Meta Adaptation for 3D Human Motion Prediction
Xiaoning Sun, Huaijiang Sun, Bin Li, Dong Wei, Weiqing Li, Jianfeng Lu; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 1042-1051
In the academic field the research on human motion prediction tasks mainly focus

es on exploiting the observed information to forecast human movements accurately in the near future horizon. However a significant gap appears when it comes to the application field as current models are all trained offline with fixed parameters that are inherently suboptimal to handle the complex yet ever-changing nature of human behaviors. To bridge this gap in this paper we introduce the task of online meta adaptation for human motion prediction based on the insight that finding "smart weights" capable of swift adjustments to suit different motion contexts along the time is a key to improving predictive accuracy. We propose MoML which ingeniously borrows the bilevel optimization spirit of model-agnostic meta-learning to transform previous predictive mistakes into strong inductive biases to guide online adaptation. This is achieved by our MoAdapter blocks that can learn error information by facilitating efficient adaptation via a few gradient steps which fine-tunes our meta-learned "smart" initialization produced by the generic predictor. Considering real-time requirements in practice we further propose Fast-MoML a more efficient variant of MoML that features a closed-form solution instead of conventional gradient update. Experimental results show that our approach can effectively bring many existing offline motion prediction models online and improves their predictive accuracy.
****

CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model
Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, An-An Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8372-8382
Generative Adversarial Networks (GANs) dominate the research field in image-based virtual try-on but have not resolved problems such as unnatural deformation of garments and the blurry generation quality. While the generative quality of diffusion models is impressive achieving controllability poses a significant challenge when applying it to virtual try-on and multiple denoising iterations limit its potential for real-time applications. In this paper we propose Controllable Accelerated virtual Try-on with Diffusion Model (CAT-DM). To enhance the controllability a basic diffusion-based virtual try-on network is designed which utilizes ControlNet to introduce additional control conditions and improves the feature extraction of garment images. In terms of acceleration CAT-DM initiates a reverse denoising process with an implicit distribution generated by a pre-trained GAN-based model. Compared with previous try-on methods based on diffusion models CAT-DM not only retains the pattern and texture details of the in-shop garment but also reduces the sampling steps without compromising generation quality. Extensive experiments demonstrate the superiority of CAT-DM against both GAN-based and diffusion-based methods in producing more realistic images and accurately reproducing garment patterns.
****

Hyperspherical Classification with Dynamic Label-to-Prototype Assignment
Mohammad Saeed Ebrahimi Saadabadi, Ali Dabouei, Sahar Rahimi Malakshan, Nasser M. Nasrabadi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17333-17342
Aiming to enhance the utilization of metric space by the parametric softmax classifier recent studies suggest replacing it with a non-parametric alternative. Although a non-parametric classifier may provide better metric space utilization it introduces the challenge of capturing inter-class relationships. A shared characteristic among prior non-parametric classifiers is the static assignment of labels to prototypes during the training i.e. each prototype consistently represents a class throughout the training course. Orthogonal to previous works we present a simple yet effective method to optimize the category assigned to each prototype (label-to-prototype assignment) during the training. To this aim we formalize the problem as a two-step optimization objective over network parameters and label-to-prototype assignment mapping. We solve this optimization using a sequential combination of gradient descent and Bipartide matching. We demonstrate the benefits of the proposed approach by conducting experiments on balanced and long-tail classification problems using different backbone network architectures. In particular our method outperforms its competitors by 1.22% accuracy on CIFAR-10

0 and 2.15% on ImageNet-200 using a metric space dimension half of the size of its competitors. \href https://github.com/msed-Ebrahimi/DL2PA_CVPR24 Code
****
VTimeLLM: Empower LLM to Grasp Video Moments
Bin Huang, Xin Wang, Hong Chen, Zihan Song, Wenwu Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14271-14280
Large language models (LLMs) have shown remarkable text understanding capabilities which have been extended as Video LLMs to handle video data for comprehending visual details. However existing Video LLMs can only provide a coarse description of the entire video failing to capture the precise start and end time boundary of specific events. In this paper we solve this issue via proposing VTimeLLM a novel Video LLM designed for fine-grained video moment understanding and reasoning with respect to time boundary. Specifically our VTimeLLM adopts a boundary-aware three-stage training strategy which respectively utilizes image-text pairs for feature alignment multiple-event videos to increase temporal-boundary awareness and high-quality video-instruction tuning to further improve temporal understanding ability as well as align with human intents. Extensive experiments demonstrate that in fine-grained time-related comprehension tasks for videos such as Temporal Video Grounding and Dense Video Captioning VTimeLLM significantly outperforms existing Video LLMs. Besides benefits from the fine-grained temporal understanding of the videos further enable VTimeLLM to beat existing Video LLMs in video dialogue benchmark showing its superior cross-modal understanding and reasoning abilities.
****
FLHetBench: Benchmarking Device and State Heterogeneity in Federated Learning
Junyuan Zhang, Shuang Zeng, Miao Zhang, Runxi Wang, Feifei Wang, Yuyin Zhou, Paul Pu Liang, Liangqiong Qu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12098-12108
Federated learning (FL) is a powerful technology that enables collaborative training of machine learning models without sharing private data among clients. The fundamental challenge in FL lies in learning over extremely heterogeneous data distributions device capacities and device state availabilities all of which adversely impact performance and communication efficiency. While data heterogeneity has been well-studied in the literature this paper introduces FLHetBench the first FL benchmark targeted toward understanding device and state heterogeneity. FLHetBench comprises two new sampling methods to generate real-world device and state databases with varying heterogeneity and new metrics for quantifying the success of FL methods under these real-world constraints. Using FLHetBench we conduct a comprehensive evaluation of existing methods and find that they struggle under these settings which inspires us to propose BiasPrompt+ a new method employing staleness-aware aggregation and fast weights to tackle these new heterogeneity challenges. Experiments on various FL tasks and datasets validate the effectiveness of our BiasPrompt+ method and highlight the value of FLHetBench in fostering the development of more efficient and robust FL solutions under real-world device and state constraints.
****
Flattening the Parent Bias: Hierarchical Semantic Segmentation in the Poincare Ball
Simon Weber, Bar?? Zöngür, Nikita Araslanov, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28223-28232
Hierarchy is a natural representation of semantic taxonomies including the ones routinely used in image segmentation. Indeed recent work on semantic segmentation reports improved accuracy from supervised training leveraging hierarchical label structures. Encouraged by these results we revisit the fundamental assumptions behind that work. We postulate and then empirically verify that the reasons for the observed improvement in segmentation accuracy may be entirely unrelated to the use of the semantic hierarchy. To demonstrate this we design a range of cross-domain experiments with a representative hierarchical approach. We find that

on the new testing domains a flat (non-hierarchical) segmentation network in which the parents are inferred from the children has superior segmentation accuracy to the hierarchical approach across the board. Complementing these findings and inspired by the intrinsic properties of hyperbolic spaces we study a more principled approach to hierarchical segmentation using the Poincare ball model. The hyperbolic representation largely outperforms the previous (Euclidean) hierarchical approach as well and is on par with our flat Euclidean baseline in terms of segmentation accuracy. However it additionally exhibits surprisingly strong calibration quality of the parent nodes in the semantic hierarchy especially on the more challenging domains. Our combined analysis suggests that the established practice of hierarchical segmentation may be limited to in-domain settings whereas flat classifiers generalize substantially better especially if they are modeled in the hyperbolic space.
****
Privacy-Preserving Optics for Enhancing Protection in Face De-Identification
Jhon Lopez, Carlos Hinojosa, Henry Arguello, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12120-12129
The modern surge in camera usage alongside widespread computer vision technology applications poses significant privacy and security concerns. Current artificial intelligence (AI) technologies aid in recognizing relevant events and assisting in daily tasks in homes offices hospitals etc. The need to access or process personal information for these purposes raises privacy concerns. While software-level solutions like face de-identification provide a good privacy/utility trade-off they present vulnerabilities to sniffing attacks. In this paper we propose a hardware-level face de-identification method to solve this vulnerability. Specifically our approach first learns an optical encoder along with a regression model to obtain a face heatmap while hiding the face identity from the source image. We also propose an anonymization framework that generates a new face using the privacy-preserving image face heatmap and a reference face image from a public dataset as input. We validate our approach with extensive simulations and hardware experiments.
****
SmartRefine: A Scenario-Adaptive Refinement Framework for Efficient Motion Prediction
Yang Zhou, Hao Shao, Letian Wang, Steven L. Waslander, Hongsheng Li, Yu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15281-15290
Predicting the future motion of surrounding agents is essential for autonomous vehicles (AVs) to operate safely in dynamic human-robot-mixed environments. Context information such as road maps and surrounding agents' states provides crucial geometric and semantic information for motion behavior prediction. To this end recent works explore two-stage prediction frameworks where coarse trajectories are first proposed and then used to select critical context information for trajectory refinement. However they either incur a large amount of computation or bring limited improvement if not both. In this paper we introduce a novel scenario-adaptive refinement strategy named SmartRefine to refine prediction with minimal additional computation. Specifically SmartRefine can comprehensively adapt refinement configurations based on each scenario's properties and smartly chooses the number of refinement iterations by introducing a quality score to measure the prediction quality and remaining refinement potential of each scenario. SmartRefine is designed as a generic and flexible approach that can be seamlessly integrated into most state-of-the-art motion prediction models. Experiments on Argoverse (1 & 2) show that our method consistently improves the prediction accuracy of multiple state-of-the-art prediction models. Specifically by adding SmartRefine to QCNet we outperform all published ensemble-free works on the Argoverse 2 leaderboard (single agent track) at submission. Comprehensive studies are also conducted to ablate design choices and explore the mechanism behind multi-iteration refinement. Codes are available at https://github.com/opendilab/SmartRefine/.
****

MVBench: A Comprehensive Multi-modal Video Understanding Benchmark
Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22195-22206
With the rapid development of Multi-modal Large Language Models (MLLMs) a number of diagnostic benchmarks have recently emerged to evaluate the comprehension capabilities of these models. However most benchmarks predominantly assess spatial understanding in the static image tasks while overlooking temporal understanding in the dynamic video tasks. To alleviate this issue we introduce a comprehensive Multi-modal Video understanding Benchmark namely MVBench which covers 20 challenging video tasks that cannot be effectively solved with a single frame. Specifically we first introduce a novel static-to-dynamic method to define these temporal-related tasks. By transforming various static tasks into dynamic ones we enable the systematic generation of video tasks that require a broad spectrum of temporal skills ranging from perception to cognition. Then guided by the task definition we automatically convert public video annotations into multiple-choice QA to evaluate each task. On one hand such a distinct paradigm allows us to build MVBench efficiently without much manual intervention. On the other hand it guarantees evaluation fairness with ground-truth video annotations avoiding the biased scoring of LLMs. Moreover we further develop a robust video MLLM baseline i.e. VideoChat2 by progressive multi-modal training with diverse instruction-tuning data. The extensive results on our MVBench reveal that the existing MLLMs are far from satisfactory in temporal understanding while our VideoChat2 largely surpasses these leading models by over 15% on MVBench.
****
Multi-Scale Video Anomaly Detection by Multi-Grained Spatio-Temporal Representation Learning
Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, Jianxin Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17385-17394
ecent progress in video anomaly detection suggests that the features of appearance and motion play crucial roles in distinguishing abnormal patterns from normal ones. However we note that the effect of spatial scales of anomalies is ignored. The fact that many abnormal events occur in limited localized regions and severe background noise interferes with the learning of anomalous changes. Meanwhile most existing methods are limited by coarse-grained modeling approaches which are inadequate for learning highly discriminative features to discriminate subtle differences between small-scale anomalies and normal patterns. To this end this paper address multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. We utilize video continuity to design three proxy tasks to perform feature learning at both coarse-grained and fine-grained levels i.e. continuity judgment discontinuity localization and missing frame estimation. In particular we formulate missing frame estimation as a contrastive learning task in feature space instead of a reconstruction task in RGB space to learn highly discriminative features. Experiments show that our proposed method outperforms state-of-the-art methods on four datasets especially in scenes with small-scale anomalies.
****
An Aggregation-Free Federated Learning for Tackling Data Heterogeneity
Yuan Wang, Huazhu Fu, Renuga Kanagavelu, Qingsong Wei, Yong Liu, Rick Siow Mong Goh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26233-26242
The performance of Federated Learning (FL) hinges on the effectiveness of utilizing knowledge from distributed datasets. Traditional FL methods adopt an aggregate-then-adapt framework where clients update local models based on a global model aggregated by the server from the previous training round. This process can cause client drift especially with significant cross-client data heterogeneity impacting model performance and convergence of the FL algorithm. To address these challenges we introduce FedAF a novel aggregation-free FL algorithm. In this framework clients collaboratively learn condensed data by leveraging peer knowledge

the server subsequently trains the global model using the condensed data and sof
t labels received from the clients. FedAF inherently avoids the issue of client
drift enhances the quality of condensed data amid notable data heterogeneity and
 improves the global model performance. Extensive numerical studies on several p
opular benchmark datasets show FedAF surpasses various state-of-the-art FL algor
ithms in handling label-skew and feature-skew data heterogeneity leading to supe
rior global model accuracy and faster convergence.
****
Generative Multimodal Models are In-Context Learners
Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming
 Rao, Jingjing Liu, Tiejun Huang, Xinlong Wang; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14398-14409
Humans can easily solve multimodal tasks in context with only a few demonstratio
ns or simple instructions which current multimodal systems largely struggle to i
mitate. In this work we demonstrate that by effectively scaling up generative mu
ltimodal models their task-agnostic in-context learning capabilities can be sign
ificantly enhanced. We introduce Emu2 a generative multimodal model with 37 bill
ion parameters which serves as a base model and general-purpose interface for a
variety of multimodal tasks. Emu2 not only achieves strong performance in few-sh
ot setting but can also be instruct-tuned to follow specific instructions such a
s visual question answering and object-grounded image generation. Emu2 even emer
ges to solve tasks that require on-the-fly reasoning such as visual prompting wh
ich existing models are unlikely to handle. We identify additional tasks where E
mu2's in-context learning can further improve and discuss its broader societal i
mpact. Our code and models will be made publicly available to facilitate future
research.
****
Synergistic Global-space Camera and Human Reconstruction from Videos
Yizhou Zhao, Tuanfeng Yang Wang, Bhiksha Raj, Min Xu, Jimei Yang, Chun-Hao Paul
Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2024, pp. 1216-1226
Remarkable strides have been made in reconstructing static scenes or human bodie
s from monocular videos. Yet the two problems have largely been approached indep
endently without much synergy. Most visual SLAM methods can only reconstruct cam
era trajectories and scene structures up to scale while most HMR methods reconst
ruct human meshes in metric scale but fall short in reasoning with cameras and s
cenes. This work introduces Synergistic Camera and Human Reconstruction (SynCHMR
) to marry the best of both worlds. Specifically we design Human-aware Metric SL
AM to reconstruct metric-scale camera poses and scene point clouds using camera-
frame HMR as a strong prior addressing depth scale and dynamic ambiguities. Cond
itioning on the dense scene recovered we further learn a Scene-aware SMPL Denois
er to enhance world-frame HMR by incorporating spatiotemporal coherency and dyna
mic scene constraints. Together they lead to consistent reconstructions of camer
a trajectories human meshes and dense scene point clouds in a common world frame
.
****
Hierarchical Intra-modal Correlation Learning for Label-free 3D Semantic Segment
ation
Xin Kang, Lei Chu, Jiahao Li, Xuejin Chen, Yan Lu; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28244-282
53
Recent methods for label-free 3D semantic segmentation aim to assist 3D model tr
aining by leveraging the open-world recognition ability of pre-trained vision la
nguage models. However these methods usually suffer from inconsistent and noisy
pseudo-labels provided by the vision language models. To address this issue we p
resent a hierarchical intra-modal correlation learning framework that captures v
isual and geometric correlations in 3D scenes at three levels: intra-set intra-s
cene and inter-scene to help learn more compact 3D representations. We refine ps
eudo-labels using intra-set correlations within each geometric consistency set a
nd align features of visually and geometrically similar points using intra-scene

and inter-scene correlation learning. We also introduce a feedback mechanism to distill the correlation learning capability into the 3D model. Experiments on both indoor and outdoor datasets show the superiority of our method. We achieve a state-of-the-art 36.6% mIoU on the ScanNet dataset and a 23.0% mIoU on the nuScenes dataset with improvements of 7.8% mIoU and 2.2% mIoU compared with previous SOTA. We also provide theoretical analysis and qualitative visualization results to discuss the mechanism and conduct thorough ablation studies to support the effectiveness of our framework.
****
Feature Re-Embedding: Towards Foundation Model-Level Performance in Computational Pathology
Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, Bo Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11343-11352
Multiple instance learning (MIL) is the most widely used framework in computational pathology encompassing sub-typing diagnosis prognosis and more. However the existing MIL paradigm typically requires an offline instance feature extractor such as a pre-trained ResNet or a foundation model. This approach lacks the capability for feature fine-tuning within the specific downstream tasks limiting its adaptability and performance. To address this issue we propose a Re-embedded Regional Transformer (RRT) for re-embedding the instance features online which captures fine-grained local features and establishes connections across different regions. Unlike existing works that focus on pre-training powerful feature extractor or designing sophisticated instance aggregator RRT is tailored to re-embed instance features online. It serves as a portable module that can seamlessly integrate into mainstream MIL models. Extensive experimental results on common computational pathology tasks validate that: 1) feature re-embedding improves the performance of MIL models based on ResNet-50 features to the level of foundation model features and further enhances the performance of foundation model features; 2) the RRT can introduce more significant performance improvements to various MIL models; 3) RRT-MIL as an RRT-enhanced AB-MIL outperforms other latest methods by a large margin. The code is available at: https://github.com/DearCaat/RRT-MIL.
****
DiffSal: Joint Audio and Video Learning for Diffusion Saliency Prediction
Junwen Xiong, Peng Zhang, Tao You, Chuanyue Li, Wei Huang, Yufei Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27273-27283
Audio-visual saliency prediction can draw support from diverse modality complements but further performance enhancement is still challenged by customized architectures as well as task-specific loss functions. In recent studies denoising diffusion models have shown more promising in unifying task frameworks owing to their inherent ability of generalization. Following this motivation a novel Diffusion architecture for generalized audio-visual Saliency prediction (DiffSal) is proposed in this work which formulates the prediction problem as a conditional generative task of the saliency map by utilizing input audio and video as the conditions. Based on the spatio-temporal audio-visual features an extra network Saliency-UNet is designed to perform multi-modal attention modulation for progressive refinement of the ground-truth saliency map from the noisy map. Extensive experiments demonstrate that the proposed DiffSal can achieve excellent performance across six challenging audio-visual benchmarks with an average relative improvement of 6.3% over the previous state-of-the-art results by six metrics.
****
Revisiting Single Image Reflection Removal In the Wild
Yurui Zhu, Xueyang Fu, Peng-Tao Jiang, Hao Zhang, Qibin Sun, Jinwei Chen, Zheng-Jun Zha, Bo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25468-25478
This research focuses on the issue of single-image reflection removal (SIRR) in real-world conditions examining it from two angles: the collection pipeline of real reflection pairs and the perception of real reflection locations. We devise an advanced reflection collection pipeline that is highly adaptable to a wide ra

nge of real-world reflection scenarios and incurs reduced costs in collecting la
rge-scale aligned reflection pairs. In the process we develop a large-scale high
-quality reflection dataset named Reflection Removal in the Wild (RRW). RRW cont
ains over 14950 high-resolution real-world reflection pairs a dataset forty-five
 times larger than its predecessors. Regarding perception of reflection location
s we identify that numerous virtual reflection objects visible in reflection ima
ges are not present in the corresponding ground-truth images. This observation d
rawn from the aligned pairs leads us to conceive the Maximum Reflection Filter (
MaxRF). The MaxRF could accurately and explicitly characterize reflection locati
ons from pairs of images. Building upon this we design a reflection location-awa
re cascaded framework specifically tailored for SIRR. Powered by these innovativ
e techniques our solution achieves superior performance than current leading met
hods across multiple real-world benchmarks. Codes and datasets are available at
\href https://github.com/zhuyr97/Reflection_RemoVal_CVPR2024 \color blue here
.
****
3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation
Zidu Wang, Xiangyu Zhu, Tianshuo Zhang, Baiqin Wang, Zhen Lei; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 1672-1682
3D Morphable Models (3DMMs) provide promising 3D face reconstructions in various
 applications. However existing methods struggle to reconstruct faces with extre
me expressions due to deficiencies in supervisory signals such as sparse or inac
curate landmarks. Segmentation information contains effective geometric contexts
 for face reconstruction. Certain attempts intuitively depend on differentiable
renderers to compare the rendered silhouettes of reconstruction with segmentatio
n which is prone to issues like local optima and gradient instability. In this p
aper we fully utilize the facial part segmentation geometry by introducing Part
Re-projection Distance Loss (PRDL). Specifically PRDL transforms facial part seg
mentation into 2D points and re-projects the reconstruction onto the image plane
. Subsequently by introducing grid anchors and computing different statistical d
istances from these anchors to the point sets PRDL establishes geometry descript
ors to optimize the distribution of the point sets for face reconstruction. PRDL
 exhibits a clear gradient compared to the renderer-based methods and presents s
tate-of-the-art reconstruction performance in extensive quantitative and qualita
tive experiments. Our project is available at https://github.com/wang-zidu/3DDFA
-V3.
****
FreeU: Free Lunch in Diffusion U-Net
Chenyang Si, Ziqi Huang, Yuming Jiang, Ziwei Liu; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4733-4743
In this paper we uncover the untapped potential of diffusion U-Net which serves
as a "free lunch" that substantially improves the generation quality on the fly.
 We initially investigate the key contributions of the U-Net architecture to the
 denoising process and identify that its main backbone primarily contributes to
denoising whereas its skip connections mainly introduce high-frequency features
into the decoder module causing the potential neglect of crucial functions intri
nsic to the backbone network. Capitalizing on this discovery we propose a simple
 yet effective method termed "FreeU" which enhances generation quality without a
dditional training or finetuning. Our key insight is to strategically re-weight
the contributions sourced from the U-Net's skip connections and backbone feature
 maps to leverage the strengths of both components of the U-Net architecture. Pr
omising results on image and video generation tasks demonstrate that our FreeU c
an be readily integrated to existing diffusion models e.g. Stable Diffusion Drea
mBooth and ControlNet to improve the generation quality with only a few lines of
 code. All you need is to adjust two scaling factors during inference.
****
Text Prompt with Normality Guidance for Weakly Supervised Video Anomaly Detectio
n
Zhiwei Yang, Jing Liu, Peng Wu; Proceedings of the IEEE/CVF Conference on Comput

er Vision and Pattern Recognition (CVPR), 2024, pp. 18899-18908
Weakly supervised video anomaly detection (WSVAD) is a challenging task. Generat
ing fine-grained pseudo-labels based on weak-label and then self-training a clas
sifier is currently a promising solution. However since the existing methods use
 only RGB visual modality and the utilization of category text information is ne
glected thus limiting the generation of more accurate pseudo-labels and affectin
g the performance of self-training. Inspired by the manual labeling process base
d on the event description in this paper we propose a novel pseudo-label generat
ion and self-training framework based on Text Prompt with Normality Guidance (TP
WNG) for WSVAD. Our idea is to transfer the rich language-visual knowledge of th
e contrastive language-image pre-training (CLIP) model for aligning the video ev
ent description text and corresponding video frames to generate pseudo-labels. S
pecifically We first fine-tune the CLIP for domain adaptation by designing two r
anking losses and a distributional inconsistency loss. Further we propose a lear
nable text prompt mechanism with the assist of a normality visual prompt to furt
her improve the matching accuracy of video event description text and video fram
es. Then we design a pseudo-label generation module based on the normality guida
nce to infer reliable frame-level pseudo-labels. Finally we introduce a temporal
 context self-adaptive learning module to learn the temporal dependencies of dif
ferent video events more flexibly and accurately. Extensive experiments show tha
t our method achieves state-of-the-art performance on two benchmark datasets UCF
-Crime and XD-Violence demonstrating the effectiveness of our proposed method.
****
SparseOcc: Rethinking Sparse Latent Representation for Vision-Based Semantic Occ
upancy Prediction
Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng,
Chao Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2024, pp. 15035-15044
Vision-based perception for autonomous driving requires an explicit modeling of
a 3D space where 2D latent representations are mapped and subsequent 3D operator
s are applied. However operating on dense latent spaces introduces a cubic time
and space complexity which limits scalability in terms of perception range or sp
atial resolution. Existing approaches compress the dense representation using pr
ojections like Bird's Eye View (BEV) or Tri-Perspective View (TPV). Although eff
icient these projections result in information loss especially for tasks like se
mantic occupancy prediction. To address this we propose SparseOcc an efficient o
ccupancy network inspired by sparse point cloud processing. It utilizes a lossle
ss sparse latent representation with three key innovations. Firstly a 3D sparse
diffuser performs latent completion using spatially decomposed 3D sparse convolu
tional kernels. Secondly a feature pyramid and sparse interpolation enhance scal
es with information from others. Finally the transformer head is redesigned as a
 sparse variant. SparseOcc achieves a remarkable 74.9% reduction on FLOPs over t
he dense baseline. Interestingly it also improves accuracy from 12.8% to 14.1% m
IOU which in part can be attributed to the sparse representation's ability to av
oid hallucinations on empty voxels.
****
SinSR: Diffusion-Based Image Super-Resolution in a Single Step
Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Z
iwei Liu, Yu Qiao, Alex C. Kot, Bihan Wen; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25796-25805
While super-resolution (SR) methods based on diffusion models exhibit promising
results their practical application is hindered by the substantial number of req
uired inference steps. Recent methods utilize the degraded images in the initial
 state thereby shortening the Markov chain. Nevertheless these solutions either
rely on a precise formulation of the degradation process or still necessitate a
relatively lengthy generation path (e.g. 15 iterations). To enhance inference sp
eed we propose a simple yet effective method for achieving single-step SR genera
tion named SinSR. Specifically we first derive a deterministic sampling process
from the most recent state-of-the-art (SOTA) method for accelerating diffusion-b
ased SR. This allows the mapping between the input random noise and the generate

d high-resolution image to be obtained in a reduced and acceptable number of inference steps during training. We show that this deterministic mapping can be distilled into a student model that performs SR within only one inference step. Additionally we propose a novel consistency-preserving loss to simultaneously leverage the ground-truth image during the distillation process ensuring that the performance of the student model is not solely bound by the feature manifold of the teacher model resulting in further performance improvement. Extensive experiments conducted on synthetic and real-world datasets demonstrate that the proposed method can achieve comparable or even superior performance compared to both previous SOTA methods and the teacher model in just one sampling step resulting in a remarkable up to x10 speedup for inference. Our code will be released at https://github.com/wyf0912/SinSR/.
****
Frequency Decoupling for Motion Magnification via Multi-Level Isomorphic Architecture
Fei Wang, Dan Guo, Kun Li, Zhun Zhong, Meng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18984-18994
Video Motion Magnification (VMM) aims to reveal subtle and imperceptible motion information of objects in the macroscopic world. Prior methods directly model the motion field from the Eulerian perspective by Representation Learning that separates shape and texture or Multi-domain Learning from phase fluctuations. Inspired by the frequency spectrum we observe that the low-frequency components with stable energy always possess spatial structure and less noise making them suitable for modeling the subtle motion field. To this end we present FD4MM a new paradigm of Frequency Decoupling for Motion Magnification with a Multi-level Isomorphic Architecture to capture multi-level high-frequency details and a stable low-frequency structure (motion field) in video space. Since high-frequency details and subtle motions are susceptible to information degradation due to their inherent subtlety and unavoidable external interference from noise we carefully design Sparse High/Low-pass Filters to enhance the integrity of details and motion structures and a Sparse Frequency Mixer to promote seamless recoupling. Besides we innovatively design a contrastive regularization for this task to strengthen the model's ability to discriminate irrelevant features reducing undesired motion magnification. Extensive experiments on both Real-world and Synthetic Datasets show that our FD4MM outperforms SOTA methods. Meanwhile FD4MM reduces FLOPs by 1.63xand boosts inference speed by 1.68xthan the latest method. Our code is available at https://github.com/Jiafei127/FD4MM.
****
Systematic Comparison of Semi-supervised and Self-supervised Learning for Medical Image Classification
Zhe Huang, Ruijie Jiang, Shuchin Aeron, Michael C. Hughes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22282-22293
In typical medical image classification problems labeled data is scarce while unlabeled data is more available. Semi-supervised learning and self-supervised learning are two different research directions that can improve accuracy by learning from extra unlabeled data. Recent methods from both directions have reported significant gains on traditional benchmarks. Yet past benchmarks do not focus on medical tasks and rarely compare self- and semi- methods together on an equal footing. Furthermore past benchmarks often handle hyperparameter tuning suboptimally. First they may not tune hyperparameters at all leading to underfitting. Second when tuning does occur it often unrealistically uses a labeled validation set that is much larger than the training set. Therefore currently published rankings might not always corroborate with their practical utility This study contributes a systematic evaluation of self- and semi- methods with a unified experimental protocol intended to guide a practitioner with scarce overall labeled data and a limited compute budget. We answer two key questions: Can hyperparameter tuning be effective with realistic-sized validation sets? If so when all methods are tuned well which self- or semi-supervised methods achieve the best accuracy? Ou

r study compares 13 representative semi- and self-supervised methods to strong l abeled-set-only baselines on 4 medical datasets. From 20000+ GPU hours of comput ation we provide valuable best practices to resource-constrained practitioners: hyperparameter tuning is effective and the semi-supervised method known as MixMa tch delivers the most reliable gains across 4 datasets.
****
ViewDiff: 3D-Consistent Image Generation with Text-to-Image Models
Lukas Höllein, Aljaž Boži?, Norman Müller, David Novotny, Hung-Yu Tseng, Christi an Richardt, Michael Zollhöfer, Matthias Nießner; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5043-5052
3D asset generation is getting massive amounts of attention inspired by the rece nt success on text-guided 2D content creation. Existing text-to-3D methods use p retrained text-to-image diffusion models in an optimization problem or fine-tune them on synthetic data which often results in non-photorealistic 3D objects wit hout backgrounds. In this paper we present a method that leverages pretrained te xt-to-image models as a prior and learn to generate multi-view images in a singl e denoising process from real-world data. Concretely we propose to integrate 3D volume-rendering and cross-frame-attention layers into each block of the existin g U-Net network of the text-to-image model. Moreover we design an autoregressive generation that renders more 3D-consistent images at any viewpoint. We train ou r model on real-world datasets of objects and showcase its capabilities to gener ate instances with a variety of high-quality shapes and textures in authentic su rroundings. Compared to the existing methods the results generated by our method are consistent and have favorable visual quality (-30% FID -37% KID).
****
Hyperbolic Learning with Synthetic Captions for Open-World Detection
Fanjie Kong, Yanbei Chen, Jiarui Cai, Davide Modolo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16762-1 6771
Open-world detection poses significant challenges as it requires the detection o f any object using either object class labels or free-form texts. Existing relat ed works often use large-scale manual annotated caption datasets for training wh ich are extremely expensive to collect. Instead we propose to transfer knowledge from vision-language models (VLMs) to enrich the open-vocabulary descriptions a utomatically. Specifically we bootstrap dense synthetic captions using pre-train ed VLMs to provide rich descriptions on different regions in images and incorpor ate these captions to train a novel detector that generalizes to novel concepts. To mitigate the noise caused by hallucination in synthetic captions we also pro pose a novel hyperbolic vision-language learning approach to impose a hierarchy between visual and caption embeddings. We call our detector "HyperLearner". We c onduct extensive experiments on a wide variety of open-world detection benchmark s (COCO LVIS Object Detection in the Wild RefCOCO) and our results show that our model consistently outperforms existing state-of-the-art methods such as GLIP G LIPv2 and Grounding DINO when using the same backbone.
****
Diffusion Models Without Attention
Jing Nathan Yan, Jiatao Gu, Alexander M. Rush; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8239-8249
In recent advancements in high-fidelity image generation Denoising Diffusion Pro babilistic Models (DDPMs) have emerged as a key player. However their applicatio n at high resolutions presents significant computational challenges. Current met hods such as patchifying expedite processes in UNet and Transformer architecture s but at the expense of representational capacity. Addressing this we introduce the Diffusion State Space Model (DiffuSSM) an architecture that supplants attent ion mechanisms with a more scalable state space model backbone. This approach ef fectively handles higher resolutions without resorting to global compression thu s preserving detailed image representation throughout the diffusion process. Our focus on FLOP-efficient architectures in diffusion training marks a significant step forward. Comprehensive evaluations on both ImageNet and LSUN datasets at t wo resolutions demonstrate that DiffuSSMs are on par or even outperform existing

diffusion models with attention modules in FID and Inception Score metrics whil
e significantly reducing total FLOP usage.
****
Interpretable Measures of Conceptual Similarity by Complexity-Constrained Descri
ptive Auto-Encoding
Alessandro Achille, Greg Ver Steeg, Tian Yu Liu, Matthew Trager, Carson Klingenb
erg, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2024, pp. 11062-11071
Quantifying the degree of similarity between images is a key copyright issue for
 image-based machine learning. In legal doctrine however determining the degree
of similarity between works requires subjective analysis and fact-finders (judge
s and juries) can demonstrate considerable variability in these subjective judge
ment calls. Images that are structurally similar can be deemed dissimilar wherea
s images of completely different scenes can be deemed similar enough to support
a claim of copying. We seek to define and compute a notion of "conceptual simila
rity" among images that captures high-level relations even among images that do
not share repeated elements or visually similar components. The idea is to use a
 base multi-modal model to generate "explanations" (captions) of visual data at
increasing levels of complexity. Then similarity can be measured by the length o
f the caption needed to discriminate between the two images: Two highly dissimil
ar images can be discriminated early in their description whereas conceptually d
issimilar ones will need more detail to be distinguished. We operationalize this
 definition and show that it correlates with subjective (averaged human evaluati
on) assessment and beats existing baselines on both image-to-image and text-to-t
ext similarity benchmarks. Beyond just providing a number our method also offers
 interpretability by pointing to the specific level of granularity of the descri
ption where the source data is differentiated.
****
Emotional Speech-driven 3D Body Animation via Disentangled Latent Diffusion
Kiran Chhatre, Radek Dan??ek, Nikos Athanasiou, Giorgio Becherini, Christopher P
eters, Michael J. Black, Timo Bolkart; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1942-1953
Existing methods for synthesizing 3D human gestures from speech have shown promi
sing results but they do not explicitly model the impact of emotions on the gene
rated gestures. Instead these methods directly output animations from speech wit
hout control over the expressed emotion. To address this limitation we present A
MUSE an emotional speech-driven body animation model based on latent diffusion.
Our observation is that content (i.e. gestures related to speech rhythm and word
 utterances) emotion and personal style are separable. To account for this AMUSE
 maps the driving audio to three disentangled latent vectors: one for content on
e for emotion and one for personal style. A latent diffusion model trained to ge
nerate gesture motion sequences is then conditioned on these latent vectors. Onc
e trained AMUSE synthesizes 3D human gestures directly from speech with control
over the expressed emotions and style by combining the content from the driving
speech with the emotion and style of another speech sequence. Randomly sampling
the noise of the diffusion model further generates variations of the gesture wit
h the same emotional expressivity. Qualitative quantitative and perceptual evalu
ations demonstrate that AMUSE outputs realistic gesture sequences. Compared to t
he state of the art the generated gestures are better synchronized with the spee
ch content and better represent the emotion expressed by the input speech. Our c
ode is available at amuse.is.tue.mpg.de.
****
3D Feature Tracking via Event Camera
Siqi Li, Zhikuan Zhou, Zhou Xue, Yipeng Li, Shaoyi Du, Yue Gao; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 18974-18983
This paper presents the first 3D feature tracking method with the corresponding
dataset. Our proposed method takes event streams from stereo event cameras as in
put to predict 3D trajectories of the target features with high-speed motion. To
 achieve this our method leverages a joint framework to predict the 2D feature m

otion offsets and the 3D feature spatial position simultaneously. A motion compe
nsation module is leveraged to overcome the feature deformation. A patch matchin
g module based on bi-polarity hypergraph modeling is proposed to robustly estima
te the feature spatial position. Meanwhile we collect the first 3D feature track
ing dataset with high-speed moving objects and ground truth 3D feature trajector
ies at 250 FPS named E-3DTrack which can be used as the first high-speed 3D feat
ure tracking benchmark. Our code and dataset could be found at: https://github.c
om/lisiqi19971013/E-3DTrack.
****
Retrieval-Augmented Layout Transformer for Content-Aware Layout Generation
Daichi Horita, Naoto Inoue, Kotaro Kikuchi, Kota Yamaguchi, Kiyoharu Aizawa; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2024, pp. 67-76
Content-aware graphic layout generation aims to automatically arrange visual ele
ments along with a given content such as an e-commerce product image. In this pa
per we argue that the current layout generation approaches suffer from the limit
ed training data for the high-dimensional layout structure. We show that a simpl
e retrieval augmentation can significantly improve the generation quality. Our m
odel which is named Retrieval-Augmented Layout Transformer (RALF) retrieves near
est neighbor layout examples based on an input image and feeds these results int
o an autoregressive generator. Our model can apply retrieval augmentation to var
ious controllable generation tasks and yield high-quality layouts within a unifi
ed architecture. Our extensive experiments show that RALF successfully generates
 content-aware layouts in both constrained and unconstrained settings and signif
icantly outperforms the baselines.
****
MSU-4S - The Michigan State University Four Seasons Dataset
Daniel Kent, Mohammed Alyaqoub, Xiaohu Lu, Hamed Khatounabadi, Kookjin Sung, Col
e Scheller, Alexander Dalat, Asma bin Thabit, Roberto Whitley, Hayder Radha; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2024, pp. 22658-22667
Public datasets such as KITTI nuScenes and Waymo have played a key role in the r
esearch and development of autonomous vehicles and advanced driver assistance sy
stems. However many of these datasets fail to incorporate a full range of drivin
g conditions; some datasets only contain clear-weather conditions underrepresent
ing or entirely missing colder weather conditions such as snow or autumn scenes
with bright colorful foliage. In this paper we present the Michigan State Univer
sity Four Seasons (MSU-4S) Dataset which contains real-world collections of auto
nomous vehicle data from varied types of driving scenarios. These scenarios were
 recorded throughout a full range of seasons and capture clear rainy snowy and f
all weather conditions at varying times of day. MSU-4S contains more than 100000
 two- and three-dimensional frames for camera lidar and radar data as well as Gl
obal Navigation Satellite System (GNSS) wheel speed and steering data all annota
ted with weather time-of-day and time-of-year. Our data includes cluttered scene
s that have large numbers of vehicles and pedestrians; and it also captures indu
strial scenes busy traffic thoroughfare with traffic lights and numerous signs a
nd scenes with dense foliage. While providing a diverse set of scenes our data i
ncorporate an important feature: virtually every scene and its corresponding lid
ar camera and radar frames were captured in four different seasons enabling unpa
ralleled object detection analysis and testing of the domain shift problem acros
s weather conditions. In that context we present detailed analyses for 3D and 2D
 object detection showing a strong domain shift effect among MSU-4S data segment
s collected across different conditions. MSU-4S will also enable advanced multim
odal fusion research including different combinations of camera-lidar-radar fusi
on which continues to be of strong interest for the computer vision autonomous d
riving and ADAS development communities. The MSU-4S dataset is available online
at https://egr.msu.edu/waves/msu4s.
****
Improving Plasticity in Online Continual Learning via Collaborative Learning
Maorong Wang, Nicolas Michel, Ling Xiao, Toshihiko Yamasaki; Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23460-23469
Online Continual Learning (CL) solves the problem of learning the ever-emerging new classification tasks from a continuous data stream. Unlike its offline counterpart in online CL the training data can only be seen once. Most existing online CL research regards catastrophic forgetting (i.e. model stability) as almost the only challenge. In this paper we argue that the model's capability to acquire new knowledge (i.e. model plasticity) is another challenge in online CL. While replay-based strategies have been shown to be effective in alleviating catastrophic forgetting there is a notable gap in research attention toward improving model plasticity. To this end we propose Collaborative Continual Learning (CCL) a collaborative learning based strategy to improve the model's capability in acquiring new concepts. Additionally we introduce Distillation Chain (DC) a collaborative learning scheme to boost the training of the models. We adapt CCL-DC to existing representative online CL works. Extensive experiments demonstrate that even if the learners are well-trained with state-of-the-art online CL methods our strategy can still improve model plasticity dramatically and thereby improve the overall performance by a large margin. The source code of our work is available at https://github.com/maorong-wang/CCL-DC.
****
InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning
Jing Shi, Wei Xiong, Zhe Lin, Hyun Joon Jung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8543-8552
Recent advances in personalized image generation have enabled pre-trained text-to-image models to learn new concepts from specific image sets. However these methods often necessitate extensive test-time finetuning for each new concept leading to inefficiencies in both time and scalability. To address this challenge we introduce InstantBooth an innovative approach leveraging existing text-to-image models for instantaneous text-guided image personalization eliminating the need for test-time finetuning. This efficiency is achieved through two primary innovations. Firstly we utilize an image encoder that transforms input images into a global embedding to grasp the general concept. Secondly we integrate new adapter layers into the pre-trained model enhancing its ability to capture intricate identity details while maintaining language coherence. Significantly our model is trained exclusively on text-image pairs without reliance on concept-specific paired images. When benchmarked against existing finetuning-based personalization techniques like DreamBooth and Textual-Inversion InstantBooth not only shows comparable proficiency in aligning language with image maintaining image quality and preserving identity but also boasts a 100-fold increase in processing speed.
****
MaxQ: Multi-Axis Query for N:M Sparsity Network
Jingyang Xiang, Siqi Li, Junhao Chen, Zhuangzhi Chen, Tianxin Huang, Linpeng Peng, Yong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15845-15854
N:M sparsity has received increasing attention due to its remarkable performance and latency trade-off compared with structured and unstructured sparsity. However existing N:M sparsity methods do not differentiate the relative importance of weights among blocks and leave important weights underappreciated. Besides they directly apply N:M sparsity to the whole network which will cause severe information loss. Thus they are still sub-optimal. In this paper we propose an efficient and effective Multi-Axis Query methodology dubbed as MaxQ to rectify these problems. During the training MaxQ employs a dynamic approach to generate soft N:M masks considering the weight importance across multiple axes. This method enhances the weights with more importance and ensures more effective updates. Meanwhile a sparsity strategy that gradually increases the percentage of N:M weight blocks is applied which allows the network to heal from the pruning-induced damage progressively. During the runtime the N:M soft masks can be precomputed as constants and folded into weights without causing any distortion to the sparse pattern and incurring additional computational overhead. Comprehensive experiments demonstrate that MaxQ achieves consistent improvements across diverse CNN architect

ures in various computer vision tasks including image classification object detection and instance segmentation. For ResNet50 with 1:16 sparse pattern MaxQ can achieve 74.6% top-1 accuracy on ImageNet and improve by over 2.8% over the state-of-the-art. Codes and checkpoints are available at https://github.com/JingyangXiang/MaxQ.
****
Part-aware Unified Representation of Language and Skeleton for Zero-shot Action Recognition
Anqi Zhu, Qiuhong Ke, Mingming Gong, James Bailey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18761-18770
While remarkable progress has been made on supervised skeleton-based action recognition the challenge of zero-shot recognition remains relatively unexplored. In this paper we argue that relying solely on aligning label-level semantics and global skeleton features is insufficient to effectively transfer locally consistent visual knowledge from seen to unseen classes. To address this limitation we introduce Part-aware Unified Representation between Language and Skeleton (PURLS) to explore visual-semantic alignment at both local and global scales. PURLS introduces a new prompting module and a novel partitioning module to generate aligned textual and visual representations across different levels. The former leverages a pre-trained GPT-3 to infer refined descriptions of the global and local (body-part-based and temporal-interval-based) movements from the original action labels. The latter employs an adaptive sampling strategy to group visual features from all body joint movements that are semantically relevant to a given description. Our approach is evaluated on various skeleton/language backbones and three large-scale datasets i.e. NTU-RGB+D 60 NTU-RGB+D 120 and a newly curated dataset Kinetics-skeleton 200. The results showcase the universality and superior performance of PURLS surpassing prior skeleton-based solutions and standard baselines from other domains. The source codes can be accessed at https://github.com/azzh1/PURLS.
****
SD2Event:Self-supervised Learning of Dynamic Detectors and Contextual Descriptors for Event Cameras
Yuan Gao, Yuqing Zhu, Xinjun Li, Yimin Du, Tianzhu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3055-3064
Event cameras offer many advantages over traditional frame-based cameras such as high dynamic range and low latency. Therefore event cameras are widely applied in diverse computer vision applications where event-based keypoint detection is a fundamental task. However achieving robust event-based keypoint detection remains challenging because the ground truth of event keypoints is difficult to obtain descriptors extracted by CNN usually lack discriminative ability in the presence of intense noise and fixed keypoint detectors are limited in detecting varied keypoint patterns. To address these challenges a novel event-based keypoint detection method is proposed by learning dynamic detectors and contextual descriptors in a self-supervised manner (SD2Event) including a contextual feature descriptor learning (CFDL) module and a dynamic keypoint detector learning (DKDL) module. The proposed SD2Event enjoys several merits. First the proposed CFDL module can model long-range contexts efficiently and effectively. Second the DKDL module generates dynamic keypoint detectors which can detect keypoints with diverse patterns across various event streams. Third the proposed self-supervised signals can guide the model's adaptation to event data. Extensive experimental results on three challenging benchmarks show that our proposed method significantly outperforms stateof-the-art event-based keypoint detection methods.
****
Composing Object Relations and Attributes for Image-Text Matching
Khoi Pham, Chuong Huynh, Ser-Nam Lim, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14354-14363
We study the visual semantic embedding problem for image-text matching. Most exi

sting work utilizes a tailored cross-attention mechanism to perform local alignm
ent across the two image and text modalities. This is computationally expensive
even though it is more powerful than the unimodal dual-encoder approach. This wo
rk introduces a dual-encoder image-text matching model leveraging a scene graph
to represent captions with nodes for objects and attributes interconnected by re
lational edges. Utilizing a graph attention network our model efficiently encode
s object-attribute and object-object semantic relations resulting in a robust an
d fast-performing system. Representing caption as a scene graph offers the abili
ty to utilize the strong relational inductive bias of graph neural networks to l
earn object-attribute and object-object relations effectively. To train the mode
l we propose losses that align the image and caption both at the holistic level
(image-caption) and the local level (image-object entity) which we show is key t
o the success of the model. Our model is termed Composition model for Object Rel
ations and Attributes CORA. Experimental results on two prominent image-text ret
rieval benchmarks Flickr30K and MS-COCO demonstrate that CORA outperforms existi
ng state-of-the-art computationally expensive cross-attention methods regarding
recall score while achieving fast computation speed of the dual encoder. Our cod
e is available at https://github.com/vkhoi/cora_cvpr24
****
Previously on ... From Recaps to Story Summarization
Aditya Kumar Singh, Dhruv Srivastava, Makarand Tapaswi; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1363
5-13646
We introduce multimodal story summarization by leveraging TV episode recaps - sh
ort video sequences interweaving key story moments from previous episodes to bri
ng viewers up to speed. We propose PlotSnap a dataset featuring two crime thrill
er TV shows with rich recaps and long episodes of 40 minutes. Story summarizatio
n labels are unlocked by matching recap shots to corresponding sub-stories in th
e episode. We propose a hierarchical model TaleSumm that processes entire episod
es by creating compact shot and dialog representations and predicts importance s
cores for each video shot and dialog utterance by enabling interactions between
local story groups. Unlike traditional summarization our method extracts multipl
e plot points from long videos. We present a thorough evaluation on story summar
ization including promising cross-series generalization. TaleSumm also shows goo
d results on classic video summarization benchmarks.
****
PaReNeRF: Toward Fast Large-scale Dynamic NeRF with Patch-based Reference
Xiao Tang, Min Yang, Penghui Sun, Hui Li, Yuchao Dai, Feng Zhu, Hojae Lee; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 5428-5438
With photo-realistic image generation Neural Radiance Field (NeRF) is widely use
d for large-scale dynamic scene reconstruction as autonomous driving simulator.
However large-scale scene reconstruction still suffers from extremely long train
ing time and rendering time. Low-resolution (LR) rendering combined with upsampl
ing can alleviate this problem but it degrades image quality. In this paper we d
esign a lightweight reference decoder which exploits prior information from know
n views to improve image reconstruction quality of new views. In addition to spe
ed up prior information search we propose an optical flow and structural similar
ity based prior information search method. Results on KITTI and VKITTI2 datasets
 show that our method significantly outperforms the baseline method in terms of
training speed rendering speed and rendering quality.
****
mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Colla
boration
Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zh
ang, Fei Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2024, pp. 13040-13051
Multi-modal Large Language Models (MLLMs) have demonstrated impressive instructi
on abilities across various open-ended tasks. However previous methods have prim
arily focused on enhancing multi-modal capabilities. In this work we introduce a

versatile multi-modal large language model mPLUG-Owl2 which effectively leverages modality collaboration to improve performance in both text and multi-modal tasks. mPLUG-Owl2 utilizes a modularized network design with the language decoder acting as a universal interface for managing different modalities. Specifically mPLUG-Owl2 incorporates shared functional modules to facilitate modality collaboration and introduces a modality-adaptive module that preserves modality-specific features. Extensive experiments reveal that mPLUG-Owl2 is capable of generalizing both text tasks and multi-modal tasks while achieving state-of-the-art performances with a single generalized model. Notably mPLUG-Owl2 is the first MLLM model that demonstrates the modality collaboration phenomenon in both pure-text and multi-modal scenarios setting a pioneering path in the development of future multi-modal foundation models.
****
Spectral and Polarization Vision: Spectro-polarimetric Real-world Dataset
Yujin Jeon, Eunsue Choi, Youngchan Kim, Yunseong Moon, Khalid Omer, Felix Heide, Seung-Hwan Baek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22098-22108
Image datasets are essential not only in validating existing methods in computer vision but also in developing new methods. Many image datasets exist consisting of trichromatic intensity images taken with RGB cameras which are designed to replicate human vision. However polarization and spectrum the wave properties of light that animals in harsh environments and with limited brain capacity often rely on remain underrepresented in existing datasets. Although there are previous spectro-polarimetric datasets they have insufficient object diversity limited illumination conditions linear-only polarization data and inadequate image count. Here we introduce two spectro-polarimetric datasets consisting of trichromatic Stokes images and hyperspectral Stokes images. These datasets encompass both linear and circular polarization; they introduce multiple spectral channels; and they feature a broad selection of real-world scenes. With our dataset in hand we analyze the spectro-polarimetric image statistics develop efficient representations of such high-dimensional data and evaluate spectral dependency of shape-from-polarization methods. As such the proposed dataset promises a foundation for data-driven spectro-polarimetric imaging and vision research.
****
Learning by Correction: Efficient Tuning Task for Zero-Shot Generative Vision-Language Reasoning
Rongjie Li, Yu Wu, Xuming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13428-13437
Generative vision-language models (VLMs) have shown impressive performance in zero-shot vision-language tasks like image captioning and visual question answering.However improving their zero-shot reasoning typically requires second-stage instruction tuning which relies heavily on human-labeled or large language model-generated annotation incurring high labeling costs. To tackle this challenge we introduce Image-Conditioned Caption Correction (ICCC) a novel pre-training task designed to enhance VLMs' zero-shot performance without the need for labeled task-aware data. The ICCC task compels VLMs to rectify mismatches between visual and language concepts thereby enhancing instruction following and text generation conditioned on visual inputs. Leveraging language structure and a lightweight dependency parser we construct data samples of ICCC task from image-text datasets with low labeling and computation costs. Experimental results on BLIP-2 and InstructBLIP demonstrate significant improvements in zero-shot image-text generation-based VL tasks through ICCC instruction tuning.
****
Supervised Anomaly Detection for Complex Industrial Images
Aimira Baitieva, David Hurych, Victor Besnier, Olivier Bernard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17754-17762
Automating visual inspection in industrial production lines is essential for increasing product quality across various industries. Anomaly detection (AD) methods serve as robust tools for this purpose. However existing public datasets prima

rily consist of images without anomalies limiting the practical application of AD methods in production settings. To address this challenge we present (1) the Valeo Anomaly Dataset (VAD) a novel real-world industrial dataset comprising 5000 images including 2000 instances of challenging real defects across more than 20 subclasses. Acknowledging that traditional AD methods struggle with this dataset we introduce (2) Segmentation-based Anomaly Detector (SegAD). First SegAD leverages anomaly maps as well as segmentation maps to compute local statistics. Next SegAD uses these statistics and an optional supervised classifier score as input features for a Boosted Random Forest (BRF) classifier yielding the final anomaly score. Our SegAD achieves state-of-the-art performance on both VAD (+2.1% AUROC) and the VisA dataset (+0.4% AUROC). The code and the models are publicly available.
****
Open3DSG: Open-Vocabulary 3D Scene Graphs from Point Clouds with Queryable Objects and Open-Set Relationships
Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, Timo Ropinski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14183-14193
Current approaches for 3D scene graph prediction rely on labeled datasets to train models for a fixed set of known object classes and relationship categories. We present Open3DSG an alternative approach to learn 3D scene graph prediction in an open world without requiring labeled scene graph data. We co-embed the features from a 3D scene graph prediction backbone with the feature space of powerful open world 2D vision language foundation models. This enables us to predict 3D scene graphs from 3D point clouds in a zero-shot manner by querying object classes from an open vocabulary and predicting the inter-object relationships from a grounded LLM with scene graph features and queried object classes as context. Open3DSG is the first 3D point cloud method to predict not only explicit open-vocabulary object classes but also open-set relationships that are not limited to a predefined label set making it possible to express rare as well as specific objects and relationships in the predicted 3D scene graph. Our experiments show that Open3DSG is effective at predicting arbitrary object classes as well as their complex inter-object relationships describing spatial supportive semantic and comparative relationships.
****
SURE: SUrvey REcipes for building reliable and robust deep networks
Yuting Li, Yingyi Chen, Xuanlong Yu, Dexiong Chen, Xi Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17500-17510
In this paper we revisit techniques for uncertainty estimation within deep neural networks and consolidate a suite of techniques to enhance their reliability. Our investigation reveals that an integrated application of diverse techniques--spanning model regularization classifier and optimization--substantially improves the accuracy of uncertainty predictions in image classification tasks. The synergistic effect of these techniques culminates in our novel SURE approach. We rigorously evaluate SURE against the benchmark of failure prediction a critical testbed for uncertainty estimation efficacy. Our results showcase a consistently better performance than models that individually deploy each technique across various datasets and model architectures. When applied to real-world challenges such as data corruption label noise and long-tailed class distribution SURE exhibits remarkable robustness delivering results that are superior or on par with current state-of-the-art specialized methods. Particularly on Animal-10N and Food-101N for learning with noisy labels SURE achieves state-of-the-art performance without any task-specific adjustments. This work not only sets a new benchmark for robust uncertainty estimation but also paves the way for its application in diverse real-world scenarios where reliability is paramount. Our code is available at https://yutingli0606.github.io/SURE/.
****
PolarRec: Improving Radio Interferometric Data Reconstruction Using Polar Coordinates

Ruoqi Wang, Zhuoyang Chen, Jiayi Zhu, Qiong Luo, Feng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12841-12850

In radio astronomy visibility data which are measurements of wave signals from radio telescopes are transformed into images for observation of distant celestial objects. However these resultant images usually contain both real sources and artifacts due to signal sparsity and other factors. One way to obtain cleaner images is to reconstruct samples into dense forms before imaging. Unfortunately existing reconstruction methods often miss some components of visibility in frequency domain so blurred object edges and persistent artifacts remain in the images. Furthermore the computation overhead is high on irregular visibility samples due to the data skew. To address these problems we propose PolarRec a transformer-encoder-conditioned reconstruction pipeline with visibility samples converted into the polar coordinate system. This coordinate system matches the way in which radio telescopes observe a celestial area as the Earth rotates. As a result visibility samples distribute in the polar system more uniformly than in the Cartesian space. Therefore we propose to use radial distance in the loss function to help reconstruct complete visibility effectively. Also we group visibility samples by their polar angles and propose a group-based encoding scheme to improve the efficiency. Our experiments demonstrate that PolarRec markedly improves imaging results by faithfully reconstructing all frequency components in the visibility domain while significantly reducing the computation cost in visibility data encoding. The code is available at https://github.com/RapidsAtHKUST/PolarRec.

****

Affine Equivariant Networks Based on Differential Invariants

Yikang Li, Yeqing Qiu, Yuxuan Chen, Lingshen He, Zhouchen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5546-5556

Convolutional neural networks benefit from translation equivariance achieving tremendous success. Equivariant networks further extend this property to other transformation groups. However most existing methods require discretization or sampling of groups leading to increased model sizes for larger groups such as the affine group. In this paper we build affine equivariant networks based on differential invariants from the viewpoint of symmetric PDEs without discretizing or sampling the group. To address the division-by-zero issue arising from fractional differential invariants of the affine group we construct a new kind of affine invariants by normalizing polynomial relative differential invariants to replace classical differential invariants. For further flexibility we design an equivariant layer which can be directly integrated into convolutional networks of various architectures. Moreover our framework for the affine group is also applicable to its continuous subgroups. We implement equivariant networks for the scale group the rotation-scale group and the affine group. Numerical experiments demonstrate the outstanding performance of our framework across classification tasks involving transformations of these groups. Remarkably under the out-of-distribution setting our model achieves a 3.37% improvement in accuracy over the main counterpart affConv on the affNIST dataset.

****

Selectively Informative Description can Reduce Undesired Embedding Entanglements in Text-to-Image Personalization

Jimyeong Kim, Jungwon Park, Wonjong Rhee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8312-8322

In text-to-image personalization a timely and crucial challenge is the tendency of generated images overfitting to the biases present in the reference images. We initiate our study with a comprehensive categorization of the biases into background nearby-object tied-object substance (in style re-contextualization) and pose biases. These biases manifest in the generated images due to their entanglement into the subject embedding. This undesired embedding entanglement not only results in the reflection of biases from the reference images into the generated images but also notably diminishes the alignment of the generated images with the given generation prompt. To address this challenge we propose SID (Selectively

Informative Description) a text description strategy that deviates from the prevalent approach of only characterizing the subject's class identification. SID is generated utilizing multimodal GPT-4 and can be seamlessly integrated into optimization-based models. We present comprehensive experimental results along with analyses of cross-attention maps subject-alignment non-subject-disentanglement and text-alignment.
****
Summarize the Past to Predict the Future: Natural Language Descriptions of Context Boost Multimodal Object Interaction Anticipation
Razvan-George Pasca, Alexey Gavryushin, Muhammad Hamza, Yen-Ling Kuo, Kaichun Mo, Luc Van Gool, Otmar Hilliges, Xi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18286-18296
We study object interaction anticipation in egocentric videos. This task requires an understanding of the spatio-temporal context formed by past actions on objects coined "action context". We propose TransFusion a multimodal transformer-based architecture for short-term object interaction anticipation. Our method exploits the representational power of language by summarizing the action context textually after leveraging pre-trained vision-language foundation models to extract the action context from past video frames. The summarized action context and the last observed video frame are processed by the multimodal fusion module to forecast the next object interaction. Experiments on the Ego4D next active object interaction dataset show the effectiveness of our multimodal fusion model and highlight the benefits of using the power of foundation models and language-based context summaries in a task where vision may appear to suffice. Our novel approach outperforms all state-of-the-art methods on both versions of the Ego4D dataset.
****
Transfer CLIP for Generalizable Image Denoising
Jun Cheng, Dong Liang, Shan Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25974-25984
Image denoising is a fundamental task in computer vision. While prevailing deep learning-based supervised and self-supervised methods have excelled in eliminating in-distribution noise their susceptibility to out-of-distribution (OOD) noise remains a significant challenge. The recent emergence of contrastive language-image pre-training (CLIP) model has showcased exceptional capabilities in open-world image recognition and segmentation. Yet the potential for leveraging CLIP to enhance the robustness of low-level tasks remains largely unexplored. This paper uncovers that certain dense features extracted from the frozen ResNet image encoder of CLIP exhibit distortion-invariant and content-related properties which are highly desirable for generalizable denoising. Leveraging these properties we devise an asymmetrical encoder-decoder denoising network which incorporates dense features including the noisy image and its multi-scale features from the frozen ResNet encoder of CLIP into a learnable image decoder to achieve generalizable denoising. The progressive feature augmentation strategy is further proposed to mitigate feature overfitting and improve the robustness of the learnable decoder. Extensive experiments and comparisons conducted across diverse OOD noises including synthetic noise real-world sRGB noise and low-dose CT image noise demonstrate the superior generalization ability of our method.
****
Smooth Diffusion: Crafting Smooth Latent Spaces in Diffusion Models
Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, Humphrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7548-7558
Recently diffusion models have made remarkable progress in text-to-image (T2I) generation synthesizing images with high fidelity and diverse contents. Despite this advancement latent space smoothness within diffusion models remains largely unexplored. Smooth latent spaces ensure that a perturbation on an input latent corresponds to a steady change in the output image. This property proves beneficial in downstream tasks including image interpolation inversion and editing. In this work we expose the non-smoothness of diffusion latent spaces by observing no

ticeable visual fluctuations resulting from minor latent variations. To tackle this issue we propose Smooth Diffusion a new category of diffusion models that can be simultaneously high-performing and smooth. Specifically we introduce Step-wise Variation Regularization to enforce the proportion between the variations of an arbitrary input latent and that of the output image is a constant at any diffusion training step. In addition we devise an interpolation standard deviation (ISTD) metric to effectively assess the latent space smoothness of a diffusion model. Extensive quantitative and qualitative experiments demonstrate that Smooth Diffusion stands out as a more desirable solution not only in T2I generation but also across various downstream tasks. Smooth Diffusion is implemented as a plug-and-play Smooth-LoRA to work with various community models. Code is available at https://github.com/SHI-Labs/Smooth-Diffusion.
****
Towards CLIP-driven Language-free 3D Visual Grounding via 2D-3D Relational Enhancement and Consistency
Yuqi Zhang, Han Luo, Yinjie Lei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13063-13072
3D visual grounding plays a crucial role in scene understanding with extensive applications in AR/VR. Despite the significant progress made in recent methods the requirement of dense textual descriptions for each individual object which is time-consuming and costly hinders their scalability. To mitigate reliance on text annotations during training researchers have explored language-free training paradigms in the 2D field via explicit text generation or implicit feature substitution. Nevertheless unlike 2D images the complexity of spatial relations in 3D coupled with the absence of robust 3D visual language pre-trained models makes it challenging to directly transfer previous strategies. To tackle the above issues in this paper we introduce a language-free training framework for 3D visual grounding. By utilizing the visual-language joint embedding in 2D large cross-modality model as a bridge we can expediently produce the pseudo-language features by leveraging the features of 2D images which are equivalent to that of real textual descriptions. We further develop a relation injection scheme with a Neighboring Relation-aware Modeling module and a Cross-modality Relation Consistency module aiming to enhance and preserve the complex relationships between the 2D and 3D embedding space. Extensive experiments demonstrate that our proposed language-free 3D visual grounding approach can obtain promising performance across three widely used datasets --ScanRefer Nr3D and Sr3D. Our codes are available at https://github.com/xibi777/3DLFVG
****
Optimal Transport Aggregation for Visual Place Recognition
Sergio Izquierdo, Javier Civera; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17658-17668
The task of Visual Place Recognition (VPR) aims to match a query image against references from an extensive database of images from different places relying solely on visual cues. State-of-the-art pipelines focus on the aggregation of features extracted from a deep backbone in order to form a global descriptor for each image. In this context we introduce SALAD (Sinkhorn Algorithm for Locally Aggregated Descriptors) which reformulates NetVLAD's soft-assignment of local features to clusters as an optimal transport problem. In SALAD we consider both feature-to-cluster and cluster-to-feature relations and we also introduce a dustbin cluster designed to selectively discard features deemed non-informative enhancing the overall descriptor quality. Additionally we leverage and fine-tune DINOv2 as a backbone which provides enhanced description power for the local features and dramatically reduces the required training time. As a result our single-stage method not only surpasses single-stage baselines in public VPR datasets but also surpasses two-stage methods that add a re-ranking with significantly higher cost.
****
FlowIE: Efficient Image Enhancement via Rectified Flow
Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13-22

Image enhancement holds extensive applications in real-world scenarios due to complex environments and limitations of imaging devices. Conventional methods are often constrained by their tailored models resulting in diminished robustness when confronted with challenging degradation conditions. In response we propose FlowIE a simple yet highly effective flow-based image enhancement framework that estimates straight-line paths from an elementary distribution to high-quality images. Unlike previous diffusion-based methods that suffer from long-time inference FlowIE constructs a linear many-to-one transport mapping via conditioned rectified flow. The rectification straightens the trajectories of probability transfer accelerating inference by an order of magnitude. This design enables our FlowIE to fully exploit rich knowledge in the pre-trained diffusion model rendering it well-suited for various real-world applications. Moreover we devise a faster inference algorithm inspired by Lagrange's Mean Value Theorem harnessing midpoint tangent direction to optimize path estimation ultimately yielding visually superior results. Thanks to these designs our FlowIE adeptly manages a diverse range of enhancement tasks within a concise sequence of fewer than 5 steps. Our contributions are rigorously validated through comprehensive experiments on synthetic and real-world datasets unveiling the compelling efficacy and efficiency of our proposed FlowIE.

****

Aligning and Prompting Everything All at Once for Universal Visual Perception

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13193-13203

Vision foundation models have been explored recently to build general-purpose vision systems. However predominant paradigms driven by casting instance-level tasks as an object-word alignment bring heavy cross-modality interaction which is not effective in prompting object detection and visual grounding. Another line of work that focuses on pixel-level tasks often encounters a large annotation gap of things and stuff and suffers from mutual interference between foreground-object and background-class segmentation. In stark contrast to the prevailing methods we present APE a universal visual perception model for aligning and prompting everything all at once in an image to perform diverse tasks i.e. detection segmentation and grounding as an instance-level sentence-object matching paradigm. Specifically APE advances the convergence of detection and grounding by reformulating language-guided grounding as open-vocabulary detection which efficiently scales up model prompting to thousands of category vocabularies and region descriptions while maintaining the effectiveness of cross-modality fusion. To bridge the granularity gap of different pixel-level tasks APE equalizes semantic and panoptic segmentation to proxy instance learning by considering any isolated regions as individual instances. APE aligns vision and language representation on broad data with natural and challenging characteristics all at once without task-specific fine-tuning. The extensive experiments on over 160 datasets demonstrate that with only one-suit of weights APE outperforms (or is on par with) the state-of-the-art models proving that an effective yet universal perception for anything aligning and prompting is indeed feasible. Codes and trained models are released at https://github.com/shenyunhang/APE.

****

Correlation-Decoupled Knowledge Distillation for Multimodal Sentiment Analysis with Incomplete Modalities

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, Lihua Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12458-12468

Multimodal sentiment analysis (MSA) aims to understand human sentiment through multimodal data. Most MSA efforts are based on the assumption of modality completeness. However in real-world applications some practical factors cause uncertain modality missingness which drastically degrades the model's performance. To this end we propose a Correlation-decoupled Knowledge Distillation (CorrKD) framework for the MSA task under uncertain missing modalities. Specifically we present

a sample-level contrastive distillation mechanism that transfers comprehensive knowledge containing cross-sample correlations to reconstruct missing semantics. Moreover a category-guided prototype distillation mechanism is introduced to capture cross-category correlations using category prototypes to align feature distributions and generate favorable joint representations. Eventually we design a response-disentangled consistency distillation strategy to optimize the sentiment decision boundaries of the student network through response disentanglement and mutual information maximization. Comprehensive experiments on three datasets indicate that our framework can achieve favorable improvements compared with several baselines.
****
Revisiting Adversarial Training at Scale
Zeyu Wang, Xianhang Li, Hongru Zhu, Cihang Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24675-24685
The machine learning community has witnessed a drastic change in the training pipeline pivoted by those "foundation models" with unprecedented scales. However the field of adversarial training is lagging behind predominantly centered around small model sizes like ResNet-50 and tiny and low-resolution datasets like CIFAR-10. To bridge this transformation gap this paper provides a modern re-examination with adversarial training investigating its potential benefits when applied at scale. Additionally we introduce an efficient and effective training strategy to enable adversarial training with giant models and web-scale data at an affordable computing cost. We denote this newly introduced framework as AdvXL. Empirical results demonstrate that AdvXL establishes new state-of-the-art robust accuracy records under AutoAttack on ImageNet-1K. For example by training on DataComp-1B dataset our AdvXL empowers a vanilla ViT-g model to substantially surpass the previous records of l_ infinity - l_ 2 - and l_ 1 -robust accuracy by margins of 11.4% 14.2% and 12.9% respectively. This achievement posits AdvXL as a pioneering approach charting a new trajectory for the efficient training of robust visual representations at significantly larger scales. Our code is available at https://github.com/UCSC-VLAA/AdvXL.
****
Towards Fairness-Aware Adversarial Learning
Yanghao Zhang, Tianle Zhang, Ronghui Mu, Xiaowei Huang, Wenjie Ruan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24746-24755
Although adversarial training (AT) has proven effective in enhancing the model's robustness the recently revealed issue of fairness in robustness has not been well addressed i.e. the robust accuracy varies significantly among different categories. In this paper instead of uniformly evaluating the model's average class performance we delve into the issue of robust fairness by considering the worst-case distribution across various classes. We propose a novel learning paradigm named Fairness-Aware Adversarial Learning (FAAL). As a generalization of conventional AT we re-define the problem of adversarial training as a min-max-max framework to ensure both robustness and fairness of the trained model. Specifically by taking advantage of distributional robust optimization our method aims to find the worst distribution among different categories and the solution is guaranteed to obtain the upper bound performance with high probability. In particular FAAL can fine-tune an unfair robust model to be fair within only two epochs without compromising the overall clean and robust accuracies. Extensive experiments on various image datasets validate the superior performance and efficiency of the proposed FAAL compared to other state-of-the-art methods.
****
LoSh: Long-Short Text Joint Prediction Network for Referring Video Object Segmentation
Linfeng Yuan, Miaojing Shi, Zijie Yue, Qijun Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14001-14010
Referring video object segmentation (RVOS) aims to segment the target instance referred by a given text expression in a video clip. The text expression normally

contains sophisticated description of the instance's appearance action and relation with others. It is therefore rather difficult for a RVOS model to capture all these attributes correspondingly in the video; in fact the model often favours more on the action- and relation-related visual attributes of the instance. This can end up with partial or even incorrect mask prediction of the target instance. We tackle this problem by taking a subject-centric short text expression from the original long text expression. The short one retains only the appearance-related information of the target instance so that we can use it to focus the model's attention on the instance's appearance. We let the model make joint predictions using both long and short text expressions; and insert a long-short cross-attention module to interact the joint features and a long-short predictions intersection loss to regulate the joint predictions. Besides the improvement on the linguistic part we also introduce a forward-backward visual consistency loss which utilizes optical flows to warp visual features between the annotated frames and their temporal neighbors for consistency. We build our method on top of two state of the art pipelines. Extensive experiments on A2D-Sentences Refer-YouTube-VOS JHMDB-Sentences and Refer-DAVIS17 show impressive improvements of our method. Code is available here.
****
MirageRoom: 3D Scene Segmentation with 2D Pre-trained Models by Mirage Projection
Haowen Sun, Yueqi Duan, Juncheng Yan, Yifan Liu, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20237-20246
Nowadays leveraging 2D images and pre-trained models to guide 3D point cloud feature representation has shown a remarkable potential to boost the performance of 3D fundamental models. While some works rely on additional data such as 2D real-world images and their corresponding camera poses recent studies target at using point cloud exclusively by designing 3D-to-2D projection. However in the indoor scene scenario existing 3D-to-2D projection strategies suffer from severe occlusions and incoherence which fail to contain sufficient information for fine-grained point cloud segmentation task. In this paper we argue that the crux of the matter resides in the basic premise of existing projection strategies that the medium is homogeneous thereby projection rays propagate along straight lines and behind objects are occluded by front ones. Inspired by the phenomenon of mirage where the occluded objects are exposed by distorted light rays due to heterogeneous medium refraction rate we propose MirageRoom by designing parametric mirage projection with heterogeneous medium to obtain series of projected images with various distorted degrees. We further develop a masked reprojection module across 2D and 3D latent space to bridge the gap between pre-trained 2D backbone and 3D point-wise features. Both quantitative and qualitative experimental results on S3DIS and ScanNet V2 demonstrate the effectiveness of our method.
****
In2SET: Intra-Inter Similarity Exploiting Transformer for Dual-Camera Compressive Hyperspectral Imaging
Xin Wang, Lizhi Wang, Xiangtian Ma, Maoqing Zhang, Lin Zhu, Hua Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24881-24891
Dual-camera compressive hyperspectral imaging (DCCHI) offers the capability to reconstruct 3D hyperspectral image (HSI) by fusing compressive and panchromatic (PAN) image which has shown great potential for snapshot hyperspectral imaging in practice. In this paper we introduce a novel DCCHI reconstruction network intra-inter similarity exploiting Transformer (In2SET). Our key insight is to make full use of the PAN image to assist the reconstruction. To this end we propose to use the intra-similarity within the PAN image as a proxy for approximating the intra-similarity in the original HSI thereby offering an enhanced content prior for more accurate HSI reconstruction. Furthermore we propose to use the inter-similarity to align the features between HSI and PAN images thereby maintaining semantic consistency between the two modalities during the reconstruction process. By integrating In2SET into a PAN-guided deep unrolling (PGDU) framework our meth

od substantially enhances the spatial-spectral fidelity and detail of the recons
tructed images providing a more comprehensive and accurate depiction of the scen
e. Experiments conducted on both real and simulated datasets demonstrate that ou
r approach consistently outperforms existing state-of-the-art methods in terms o
f reconstruction quality and computational complexity. The code is available at
https://github.com/2JONAS/In2SET.
****
Dual Prototype Attention for Unsupervised Video Object Segmentation
Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Dogyoon Lee, Heeseung Choi, Ig-Jae Kim,
 Sangyoun Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2024, pp. 19238-19247
Unsupervised video object segmentation (VOS) aims to detect and segment the most
 salient object in videos. The primary techniques used in unsupervised VOS are 1
) the collaboration of appearance and motion information; and 2) temporal fusion
 between different frames. This paper proposes two novel prototype-based attenti
on mechanisms inter-modality attention (IMA) and inter-frame attention (IFA) to
incorporate these techniques via dense propagation across different modalities a
nd frames. IMA densely integrates context information from different modalities
based on a mutual refinement. IFA injects global context of a video to the query
 frame enabling a full utilization of useful properties from multiple frames. Ex
perimental results on public benchmark datasets demonstrate that our proposed ap
proach outperforms all existing methods by a substantial margin. The proposed tw
o components are also thoroughly validated via ablative study.
****
Look-Up Table Compression for Efficient Image Restoration
Yinglong Li, Jiacheng Li, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26016-26025
Look-Up Table (LUT) has recently gained increasing attention for restoring High-
Quality (HQ) images from Low-Quality (LQ) observations thanks to its high comput
ational efficiency achieved through a "space for time" strategy of caching learn
ed LQ-HQ pairs. However incorporating multiple LUTs for improved performance com
es at the cost of a rapidly growing storage size which is ultimately restricted
by the allocatable on-device cache size. In this work we propose a novel LUT com
pression framework to achieve a better trade-off between storage size and perfor
mance for LUT-based image restoration models. Based on the observation that most
 cached LQ image patches are distributed along the diagonal of a LUT we devise a
 Diagonal-First Compression (DFC) framework where diagonal LQ-HQ pairs are prese
rved and carefully re-indexed to maintain the representation capacity while non-
diagonal pairs are aggressively subsampled to save storage. Extensive experiment
s on representative image restoration tasks demonstrate that our DFC framework s
ignificantly reduces the storage size of LUT-based models (including our new des
ign) while maintaining their performance. For instance DFC saves up to 90% of st
orage at a negligible performance drop for x4 super-resolution. The source code
is available on GitHub: https://github.com/leenas233/DFC.
****
TextNeRF: A Novel Scene-Text Image Synthesis Method based on Neural Radiance Fie
lds
Jialei Cui, Jianwei Du, Wenzhuo Liu, Zhouhui Lian; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22272-222
81
Acquiring large-scale well-annotated datasets is essential for training robust s
cene text detectors yet the process is often resource-intensive and time-consumi
ng. While some efforts have been made to explore the synthesis of scene text ima
ges a notable gap remains between synthetic and authentic data. In this paper we
 introduce a novel method that utilizes Neural Radiance Fields (NeRF) to model r
eal-world scenes and emulate the data collection process by rendering images fro
m diverse camera perspectives enriching the variability and realism of the synth
esized data. A semi-supervised learning framework is proposed to categorize sema
ntic regions within 3D scenes ensuring consistent labeling of text regions acros
s various viewpoints. Our method also models the pose and view-dependent appeara

nce of text regions thereby offering precise control over camera poses and significantly improving the realism of text insertion and editing within scenes. Employing our technique on real-world scenes has led to the creation of a novel scene text image dataset. Compared to other existing benchmarks the proposed dataset is distinctive in providing not only standard annotations such as bounding boxes and transcriptions but also the information of 3D pose attributes for text regions enabling a more detailed evaluation of the robustness of text detection algorithms. Through extensive experiments we demonstrate the effectiveness of our proposed method in enhancing the performance of scene text detectors.
****
Dr.Hair: Reconstructing Scalp-Connected Hair Strands without Pre-Training via Differentiable Rendering of Line Segments
Yusuke Takimoto, Hikari Takehara, Hiroyuki Sato, Zihao Zhu, Bo Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20601-20611
In the film and gaming industries achieving a realistic hair appearance typically involves the use of strands originating from the scalp. However reconstructing these strands from observed surface images of hair presents significant challenges. The difficulty in acquiring Ground Truth (GT) data has led state-of-the-art learning-based methods to rely on pre-training with manually prepared synthetic CG data. This process is not only labor-intensive and costly but also introduces complications due to the domain gap when compared to real-world data. In this study we propose an optimization-based approach that eliminates the need for pre-training. Our method represents hair strands as line segments growing from the scalp and optimizes them using a novel differentiable rendering algorithm. To robustly optimize a substantial number of slender explicit geometries we introduce 3D orientation estimation utilizing global optimization strand initialization based on Laplace's equation and reparameterization that leverages geometric connectivity and spatial proximity. Unlike existing optimization-based methods our method is capable of reconstructing internal hair flow in an absolute direction. Our method exhibits robust and accurate inverse rendering surpassing the quality of existing methods and significantly improving processing speed.
****
Improving Training Efficiency of Diffusion Models via Multi-Stage Framework and Tailored Multi-Decoder Architecture
Huijie Zhang, Yifu Lu, Ismail Alkhouri, Saiprasad Ravishankar, Dogyoon Song, Qing Qu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7372-7381
Diffusion models emerging as powerful deep generative tools excel in various applications. They operate through a two-steps process: introducing noise into training samples and then employing a model to convert random noise into new samples (e.g. images). However their remarkable generative performance is hindered by slow training and sampling. This is due to the necessity of tracking extensive forward and reverse diffusion trajectories and employing a large model with numerous parameters across multiple timesteps (i.e. noise levels). To tackle these challenges we present a multi-stage framework inspired by our empirical findings. These observations indicate the advantages of employing distinct parameters tailored to each timestep while retaining universal parameters shared across all time steps. Our approach involves segmenting the time interval into multiple stages where we employ custom multi-decoder U-net architecture that blends time-dependent models with a universally shared encoder. Our framework enables the efficient distribution of computational resources and mitigates inter-stage interference which substantially improves training efficiency. Extensive numerical experiments affirm the effectiveness of our framework showcasing significant training and sampling efficiency enhancements on three state-of-the-art diffusion models including large-scale latent diffusion models. Furthermore our ablation studies illustrate the impact of two important components in our framework: (i) a novel timestep clustering algorithm for stage division and (ii) an innovative multi-decoder U-net architecture seamlessly integrating universal and customized hyperparameters.

****
In-Context Matting
He Guo, Zixuan Ye, Zhiguo Cao, Hao Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3711-3720
We introduce in-context matting a novel task setting of image matting. Given a reference image of a certain foreground and guided priors such as points scribbles and masks in-context matting enables automatic alpha estimation on a batch of target images of the same foreground category without additional auxiliary input. This setting marries good performance in auxiliary input-based matting and ease of use in automatic matting which finds a good trade-off between customization and automation. To overcome the key challenge of accurate foreground matching we introduce IconMatting an in-context matting model built upon a pre-trained text-to-image diffusion model. Conditioned on inter- and intra-similarity matching IconMatting can make full use of reference context to generate accurate target alpha mattes. To benchmark the task we also introduce a novel testing dataset ICM-57 covering 57 groups of real-world images. Quantitative and qualitative results on the ICM-57 testing set show that IconMatting rivals the accuracy of trimap-based matting while retaining the automation level akin to automatic matting. Code is available at https://github.com/tiny-smart/in-context-matting.
****
Navigate Beyond Shortcuts: Debiased Learning Through the Lens of Neural Collapse
Yining Wang, Junjie Sun, Chenyue Wang, Mi Zhang, Min Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12322-12331
Recent studies have noted an intriguing phenomenon termed Neural Collapse that is when the neural networks establish the right correlation between feature spaces and the training targets their last-layer features together with the classifier weights will collapse into a stable and symmetric structure. In this paper we extend the investigation of Neural Collapse to the biased datasets with imbalanced attributes. We observe that models will easily fall into the pitfall of shortcut learning and form a biased non-collapsed feature space at the early period of training which is hard to reverse and limits the generalization capability. To tackle the root cause of biased classification we follow the recent inspiration of prime training and propose an avoid-shortcut learning framework without additional training complexity. With well-designed shortcut primes based on Neural Collapse structure the models are encouraged to skip the pursuit of simple shortcuts and naturally capture the intrinsic correlations. Experimental results demonstrate that our method induces a better convergence property during training and achieves state-of-the-art generalization performance on both synthetic and real-world biased datasets.
****
DiVa-360: The Dynamic Visual Dataset for Immersive Neural Fields
Cheng-You Lu, Peisen Zhou, Angela Xing, Chandradeep Pokhariya, Arnab Dey, Ishaan Nikhil Shah, Rugved Mavidipalli, Dylan Hu, Andrew I. Comport, Kefan Chen, Srinath Sridhar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22466-22476
Advances in neural fields are enabling high-fidelity capture of the shape and appearance of dynamic 3D scenes. However their capabilities lag behind those offered by conventional representations such as 2D videos because of algorithmic challenges and the lack of large-scale multi-view real-world datasets. We address the dataset limitation with DiVa-360 a real-world 360? dynamic visual dataset that contains synchronized high-resolution and long-duration multi-view video sequences of table-scale scenes captured using a customized low-cost system with 53 cameras. It contains 21 object-centric sequences categorized by different motion types 25 intricate hand-object interaction sequences and 8 long-duration sequences for a total of 17.4 M image frames. In addition we provide foreground-background segmentation masks synchronized audio and text descriptions. We benchmark the state-of-the-art dynamic neural field methods on DiVa-360 and provide insights about existing methods and future challenges on long-duration neural field capture.

****
A Subspace-Constrained Tyler's Estimator and its Applications to Structure from Motion
Feng Yu, Teng Zhang, Gilad Lerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14575-14584
We present the subspace-constrained Tyler's estimator (STE) designed for recovering a low-dimensional subspace within a dataset that may be highly corrupted with outliers. STE is a fusion of the Tyler's M-estimator (TME) and a variant of the fast median subspace. Our theoretical analysis suggests that under a common inlier-outlier model STE can effectively recover the underlying subspace even when it contains a smaller fraction of inliers relative to other methods in the field of robust subspace recovery. We apply STE in the context of Structure from Motion (SfM) in two ways: for robust estimation of the fundamental matrix and for the removal of outlying cameras enhancing the robustness of the SfM pipeline. Numerical experiments confirm the state-of-the-art performance of our method in these applications. This research makes significant contributions to the field of robust subspace recovery particularly in the context of computer vision and 3D reconstruction.
****
FSC: Few-point Shape Completion
Xianzu Wu, Xianfeng Wu, Tianyu Luan, Yajing Bai, Zhongyuan Lai, Junsong Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26077-26087
While previous studies have demonstrated successful 3D object shape completion with a sufficient number of points they often fail in scenarios when a few points e.g. tens of points are observed. Surprisingly via entropy analysis we find that even a few points e.g. 64 points could retain substantial information to help recover the 3D shape of the object. To address the challenge of shape completion with very sparse point clouds we then propose Few-point Shape Completion (FSC) model which contains a novel dual-branch feature extractor for handling extremely sparse inputs coupled with an extensive branch for maximal point utilization with a saliency branch for dynamic importance assignment. This model is further bolstered by a two-stage revision network that refines both the extracted features and the decoder output enhancing the detail and authenticity of the completed point cloud. Our experiments demonstrate the feasibility of recovering 3D shapes from a few points. The proposed Few-point Shape Completion (FSC) model outperforms previous methods on both few-point inputs and many-point inputs and shows good generalizability to different object categories.
****
CAD: Photorealistic 3D Generation via Adversarial Distillation
Ziyu Wan, Despoina Paschalidou, Ian Huang, Hongyu Liu, Bokui Shen, Xiaoyu Xiang, Jing Liao, Leonidas Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10194-10207
The increased demand for 3D data in AR/VR robotics and gaming applications gave rise to powerful generative pipelines capable of synthesizing high-quality 3D objects. Most of these models rely on the Score Distillation Sampling (SDS) algorithm to optimize a 3D representation such that the rendered image maintains a high likelihood as evaluated by a pre-trained diffusion model. However this distillation process involves finding a correct mode in the high-dimensional and large-variance distribution produced by the diffusion model. This task is challenging and often leads to issues such as over-saturation over-smoothing and Janus-like artifacts in the 3D generation. In this paper we propose a novel learning paradigm for 3D synthesis that utilizes pre-trained diffusion models. Instead of focusing on mode-seeking our method directly models the distribution discrepancy between multi-view renderings and diffusion priors in an adversarial manner which unlocks the generation of high-fidelity and photorealistic 3D content conditioned on a single image and prompt. Moreover by harnessing the latent space of GANs and expressive diffusion model priors our method enables a wide variety of 3D applications including single-view reconstruction high diversity generation and continuous 3D interpolation in open domain. Our experiments demonstrate the superior

ity of our pipeline compared to previous works in terms of generation quality an
d diversity.
****
Enhancing Vision-Language Pre-training with Rich Supervisions
Yuan Gao, Kunyu Shi, Pengkai Zhu, Edouard Belval, Oren Nuriel, Srikar Appalaraju
, Shabnam Ghadar, Zhuowen Tu, Vijay Mahadevan, Stefano Soatto; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 13480-13491
We propose Strongly Supervised pre-training with ScreenShots (S4) - a novel pre-
training paradigm for Vision-Language Models using data from large-scale web scr
eenshot rendering. Using web screenshots unlocks a treasure trove of visual and
textual cues that are not present in using image-text pairs. In S4 we leverage t
he inherent tree-structured hierarchy of HTML elements and the spatial localizat
ion to carefully design 10 pre-training tasks with large scale annotated data. T
hese tasks resemble downstream tasks across different domains and the annotation
s are cheap to obtain. We demonstrate that compared to current screenshot pre-tr
aining objectives our innovative pre-training method significantly enhances perf
ormance of image-to-text model in nine varied and popular downstream tasks - up
to 76.1% improvements on Table Detection and at least 1% on Widget Captioning.
****
T-VSL: Text-Guided Visual Sound Source Localization in Mixtures
Tanvir Mahmud, Yapeng Tian, Diana Marculescu; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26742-26751
Visual sound source localization poses a significant challenge in identifying th
e semantic region of each sounding source within a video. Existing self-supervis
ed and weakly supervised source localization methods struggle to accurately dist
inguish the semantic regions of each sounding object particularly in multi-sourc
e mixtures. These methods often rely on audio-visual correspondence as guidance
which can lead to substantial performance drops in complex multi-source localiza
tion scenarios. The lack of access to individual source sounds in multi-source m
ixtures during training exacerbates the difficulty of learning effective audio-v
isual correspondence for localization. To address this limitation in this paper
we propose incorporating the text modality as an intermediate feature guide usin
g tri-modal joint embedding models (e.g. AudioCLIP) to disentangle the semantic
audio-visual source correspondence in multi-source mixtures. Our framework dubbe
d T-VSL begins by predicting the class of sounding entities in mixtures. Subsequ
ently the textual representation of each sounding source is employed as guidance
 to disentangle fine-grained audio-visual source correspondence from multi-sourc
e mixtures leveraging the tri-modal AudioCLIP embedding. This approach enables o
ur framework to handle a flexible number of sources and exhibits promising zero-
shot transferability to unseen classes during test time. Extensive experiments c
onducted on the MUSIC VGGSound and VGGSound-Instruments datasets demonstrate sig
nificant performance improvements over state-of-the-art methods. Code is release
d at https://github.com/enyac-group/T-VSL/tree/main.
****
DemoCaricature: Democratising Caricature Generation with a Rough Sketch
Dar-Yen Chen, Ayan Kumar Bhunia, Subhadeep Koley, Aneeshan Sain, Pinaki Nath Cho
wdhury, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2024, pp. 8629-8639
In this paper we democratise caricature generation empowering individuals to eff
ortlessly craft personalised caricatures with just a photo and a conceptual sket
ch. Our objective is to strike a delicate balance between abstraction and identi
ty while preserving the creativity and subjectivity inherent in a sketch. To ach
ieve this we present Explicit Rank-1 Model Editing alongside single-image person
alisation selectively applying nuanced edits to cross-attention layers for a sea
mless merge of identity and style. Additionally we propose Random Mask Reconstru
ction to enhance robustness directing the model to focus on distinctive identity
 and style features. Crucially our aim is not to replace artists but to eliminat
e accessibility barriers allowing enthusiasts to engage in the artistry.
****

CapHuman: Capture Your Moments in Parallel Universes
Chao Liang, Fan Ma, Linchao Zhu, Yingying Deng, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6400-6409
We concentrate on a novel human-centric image synthesis task that is given only one reference facial photograph it is expected to generate specific individual images with diverse head positions poses facial expressions and illuminations in different contexts. To accomplish this goal we argue that our generative model should be capable of the following favorable characteristics: (1) a strong visual and semantic understanding of our world and human society for basic object and human image generation. (2) generalizable identity preservation ability. (3) flexible and fine-grained head control. Recently large pre-trained text-to-image diffusion models have shown remarkable results serving as a powerful generative foundation. As a basis we aim to unleash the above two capabilities of the pre-trained model. In this work we present a new framework named CapHuman. We embrace the "encode then learn to align" paradigm which enables generalizable identity preservation for new individuals without cumbersome tuning at inference. CapHuman encodes identity features and then learns to align them into the latent space. Moreover we introduce the 3D facial prior to equip our model with control over the human head in a flexible and 3D-consistent manner. Extensive qualitative and quantitative analyses demonstrate our CapHuman can produce well-identity-preserved photo-realistic and high-fidelity portraits with content-rich representations and various head renditions superior to established baselines. Code and checkpoint will be released at https://github.com/VamosC/CapHuman.
****
SDPose: Tokenized Pose Estimation via Circulation-Guide Self-Distillation
Sichen Chen, Yingyi Zhang, Siming Huang, Ran Yi, Ke Fan, Ruixin Zhang, Peixian Chen, Jun Wang, Shouhong Ding, Lizhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1082-1090
Recently transformer-based methods have achieved state-of-the-art prediction quality on human pose estimation(HPE). Nonetheless most of these top-performing transformer-based models are too computation-consuming and storage-demanding to deploy on edge computing platforms. Those transformer-based models that require fewer resources are prone to under-fitting due to their smaller scale and thus perform notably worse than their larger counterparts. Given this conundrum we introduce SDPose a new self-distillation method for improving the performance of small transformer-based models. To mitigate the problem of under-fitting we design a transformer module named Multi-Cycled Transformer(MCT) based on multiple-cycled forwards to more fully exploit the potential of small model parameters. Further in order to prevent the additional inference compute-consuming brought by MCT we introduce a self-distillation scheme extracting the knowledge from the MCT module to a naive forward model. Specifically on the MSCOCO validation dataset SDPose-T obtains 69.7% mAP with 4.4M parameters and 1.8 GFLOPs. Furthermore SDPose-S-V2 obtains 73.5% mAP on the MSCOCO validation dataset with 6.2M parameters and 4.7 GFLOPs achieving a new state-of-the-art among predominant tiny neural network methods.
****
Authentic Hand Avatar from a Phone Scan via Universal Hand Model
Gyeongsik Moon, Weipeng Xu, Rohan Joshi, Chenglei Wu, Takaaki Shiratori; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2029-2038
The authentic 3D hand avatar with every identifiable information such as hand shapes and textures is necessary for immersive experiences in AR/VR. In this paper we present a universal hand model (UHM) which 1) can universally represent high-fidelity 3D hand meshes of arbitrary identities (IDs) and 2) can be adapted to each person with a short phone scan for the authentic hand avatar. For effective universal hand modeling we perform tracking and modeling at the same time while previous 3D hand models perform them separately. The conventional separate pipeline suffers from the accumulated errors from the tracking stage which cannot be recovered in the modeling stage. On the other hand ours does not suffer from th

e accumulated errors while having a much more concise overall pipeline. We additionally introduce a novel image matching loss function to address a skin sliding during the tracking and modeling while existing works have not focused on it much. Finally using learned priors from our UHM we effectively adapt our UHM to each person's short phone scan for the authentic hand avatar.
****
VCoder: Versatile Vision Encoders for Multimodal Large Language Models
Jitesh Jain, Jianwei Yang, Humphrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27992-28002
Humans possess the remarkable skill of Visual Perception the ability to see and understand the seen helping them make sense of the visual world and in turn reason. Multimodal Large Language Models (MLLM) have recently achieved impressive performance on vision-language tasks ranging from visual question-answering and image captioning to visual reasoning and image generation. However when prompted to identify or count (perceive) the entities in a given image existing MLLM systems fail. Working towards developing an accurate MLLM system for perception and reasoning we propose using Versatile vision enCoders (VCoder) as perception eyes for Multimodal LLMs. We feed the VCoder with perception modalities such as segmentation or depth maps improving the MLLM's perception abilities. Secondly we leverage the images from COCO and outputs from off-the-shelf vision perception models to create our COCO Segmentation Text (COST) dataset for training and evaluating MLLMs on the object perception task. Thirdly we introduce metrics to assess the object perception abilities in MLLMs on our COST dataset. Lastly we provide extensive experimental evidence proving the VCoder's improved object-level perception skills over existing Multimodal LLMs including GPT-4V. We open-source our dataset code and models to promote research.
****
Event-based Visible and Infrared Fusion via Multi-task Collaboration
Mengyue Geng, Lin Zhu, Lizhi Wang, Wei Zhang, Ruiqin Xiong, Yonghong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26929-26939
Visible and Infrared image Fusion (VIF) offers a comprehensive scene description by combining thermal infrared images with the rich textures from visible cameras. However conventional VIF systems may capture over/under exposure or blurry images in extreme lighting and high dynamic motion scenarios leading to degraded fusion results. To address these problems we propose a novel Event-based Visible and Infrared Fusion (EVIF) system that employs a visible event camera as an alternative to traditional frame-based cameras for the VIF task. With extremely low latency and high dynamic range event cameras can effectively address blurriness and are robust against diverse luminous ranges. To produce high-quality fused images we develop a multi-task collaborative framework that simultaneously performs event-based visible texture reconstruction event-guided infrared image deblurring and visible-infrared fusion. Rather than independently learning these tasks our framework capitalizes on their synergy leveraging cross-task event enhancement for efficient deblurring and bi-level min-max mutual information optimization to achieve higher fusion quality. Experiments on both synthetic and real data show that EVIF achieves remarkable performance in dealing with extreme lighting conditions and high-dynamic scenes ensuring high-quality fused images across a broad range of practical scenarios.
****
Open-World Semantic Segmentation Including Class Similarity
Matteo Sodano, Federico Magistri, Lucas Nunes, Jens Behley, Cyrill Stachniss; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3184-3194
Interpreting camera data is key for autonomously acting systems such as autonomous vehicles. Vision systems that operate in real-world environments must be able to understand their surroundings and need the ability to deal with novel situations. This paper tackles open-world semantic segmentation i.e. the variant of interpreting image data in which objects occur that have not been seen during training. We propose a novel approach that performs accurate closed-world semantic s

egmentation and at the same time can identify new categories without requiring any additional training data. Our approach additionally provides a similarity measure for every newly discovered class in an image to a known category which can be useful information in downstream tasks such as planning or mapping. Through extensive experiments we show that our model achieves state-of-the-art results on classes known from training data as well as for anomaly segmentation and can distinguish between different unknown classes.
****
RegionPLC: Regional Point-Language Contrastive Learning for Open-World 3D Scene Understanding
Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19823-19832
We propose a lightweight and scalable Regional Point-Language Contrastive learning framework namely RegionPLC for open-world 3D scene understanding aiming to identify and recognize open-set objects and categories. Specifically based on our empirical studies we introduce a 3D-aware SFusion strategy that fuses 3D vision-language pairs derived from multiple 2D foundation models yielding high-quality dense region-level language descriptions without human 3D annotations. Subsequently we devise a region-aware point-discriminative contrastive learning objective to enable robust and effective 3D learning from dense regional language supervision. We carry out extensive experiments on ScanNet ScanNet200 and nuScenes datasets and our model outperforms prior 3D open-world scene understanding approaches by an average of 17.2% and 9.1% for semantic and instance segmentation respectively while maintaining greater scalability and lower resource demands. Furthermore our method has the flexibility to be effortlessly integrated with language models to enable open-ended grounded 3D reasoning without extra task-specific training. Code will be released.
****
Adaptive VIO: Deep Visual-Inertial Odometry with Online Continual Learning
Youqi Pan, Wugen Zhou, Yingdian Cao, Hongbin Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18019-18028
Visual-inertial odometry (VIO) has demonstrated remarkable success due to its low-cost and complementary sensors. However existing VIO methods lack the generalization ability to adjust to different environments and sensor attributes. In this paper we propose Adaptive VIO a new monocular visual-inertial odometry that combines online continual learning with traditional nonlinear optimization. Adaptive VIO comprises two networks to predict visual correspondence and IMU bias. Unlike end-to-end approaches that use networks to fuse the features from two modalities (camera and IMU) and predict poses directly we combine neural networks with visual-inertial bundle adjustment in our VIO system. The optimized estimates will be fed back to the visual and IMU bias networks refining the networks in a self-supervised manner. Such a learning-optimization-combined framework and feedback mechanism enable the system to perform online continual learning. Experiments demonstrate that our Adaptive VIO manifests adaptive capability on EuRoC and TUM-VI datasets. The overall performance exceeds the currently known learning-based VIO methods and is comparable to the state-of-the-art optimization-based methods.
****
Towards Memorization-Free Diffusion Models
Chen Chen, Daochang Liu, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8425-8434
Pretrained diffusion models and their outputs are widely accessible due to their exceptional capacity for synthesizing high-quality images and their open-source nature. The users however may face litigation risks owing to the models' tendency to memorize and regurgitate training data during inference. To address this we introduce Anti-Memorization Guidance (AMG) a novel framework employing three targeted guidance strategies for the main causes of memorization: image and caption duplication and highly specific user prompts. Consequently AMG ensures memori

zation-free outputs while maintaining high image quality and text alignment leveraging the synergy of its guidance methods each indispensable in its own right. AMG also features an innovative automatic detection system for potential memorization during each step of inference process allows selective application of guidance strategies minimally interfering with the original sampling process to preserve output utility. We applied AMG to pretrained Denoising Diffusion Probabilistic Models (DDPM) and Stable Diffusion across various generation tasks. The results demonstrate that AMG is the first approach to successfully eradicates all instances of memorization with no or marginal impacts on image quality and text-alignment as evidenced by FID and CLIP scores.
****
Generalized Large-Scale Data Condensation via Various Backbone and Statistical Matching
Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, Zhiqiang Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16709-16718
The lightweight "local-match-global" matching introduced by SRe2L successfully creates a distilled dataset with comprehensive information on the full 224x224 ImageNet-1k. However this one-sided approach is limited to a particular backbone layer and statistics which limits the improvement of the generalization of a distilled dataset. We suggest that sufficient and various "local-match-global" matching are more precise and effective than a single one and has the ability to create a distilled dataset with richer information and better generalization. We call this perspective "generalized matching" and propose Generalized Various Backbone and Statistical Matching (G-VBSM) in this work which aims to create a synthetic dataset with densities ensuring consistency with the complete dataset across various backbones layers and statistics. As experimentally demonstrated G-VBSM is the first algorithm to obtain strong performance across both small-scale and large-scale datasets. Specifically G-VBSM achieves a performance of 38.7% on CIFAR-100 with 128-width ConvNet 47.6% on Tiny-ImageNet with ResNet18 and 31.4% on the full 224x224 ImageNet-1k with ResNet18 under images per class (IPC) 10 50 and 10 respectively. These results surpass all SOTA methods by margins of 3.9% 6.5% and 10.1% respectively.
****
Three Pillars Improving Vision Foundation Model Distillation for Lidar
Gilles Puy, Spyros Gidaris, Alexandre Boulch, Oriane Siméoni, Corentin Sautier, Patrick Pérez, Andrei Bursuc, Renaud Marlet; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21519-21529
Self-supervised image backbones can be used to address complex 2D tasks (e.g. semantic segmentation object discovery) very efficiently and with little or no downstream supervision. Ideally 3D backbones for lidar should be able to inherit these properties after distillation of these powerful 2D features. The most recent methods for image-to-lidar distillation on autonomous driving data show promising results obtained thanks to distillation methods that keep improving. Yet we still notice a large performance gap when measuring by linear probing the quality of distilled vs fully supervised features. In this work instead of focusing only on the distillation method we study the effect of three pillars for distillation: the 3D backbone the pretrained 2D backbone and the pretraining 2D+3D dataset. In particular thanks to our scalable distillation method named ScaLR we show that scaling the 2D and 3D backbones and pretraining on diverse datasets leads to a substantial improvement of the feature quality. This allows us to significantly reduce the gap between the quality of distilled and fully-supervised 3D features and to improve the robustness of the pretrained backbones to domain gaps and perturbations.
****
On Train-Test Class Overlap and Detection for Image Retrieval
Chull Hwan Song, Jooyoung Yoon, Taebaek Hwang, Shunghyun Choi, Yeong Hyeon Gu, Yannis Avrithis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17375-17384
How important is it for training and evaluation sets to not have class overlap i

n image retrieval? We revisit Google Landmarks v2 clean the most popular training set by identifying and removing class overlap with Revisited Oxford and Paris the most popular training set. By comparing the original and the new RGLDv2-clean on a benchmark of reproduced state-of-the-art methods our findings are striking. Not only is there a dramatic drop in performance but it is inconsistent across methods changing the ranking. What does it take to focus on objects or interest and ignore background clutter when indexing? Do we need to analyze the evaluation set? Do we need to train an object detector and the representation separately? Do we need location supervision? We introduce Single-stage Detect-to-Retrieve (CiDeR) an end-to-end single-stage pipeline to detect objects of interest and extract a global image representation. We outperform previous state-of-the-art on both existing training sets and the new RGLDv2-clean.
****
AttriHuman-3D: Editable 3D Human Avatar Generation with Attribute Decomposition and Indexing
Fan Yang, Tianyi Chen, Xiaosheng He, Zhongang Cai, Lei Yang, Si Wu, Guosheng Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10596-10605
Editable 3D-aware generation which supports user-interacted editing has witnessed rapid development recently. However existing editable 3D GANs either fail to achieve high-accuracy local editing or suffer from huge computational costs. We propose AttriHuman-3D an editable 3D human generation model which address the aforementioned problems with attribute decomposition and indexing. The core idea of the proposed model is to generate all attributes (e.g. human body hair clothes and so on) in an overall attribute space with six feature planes which are then decomposed and manipulated with different attribute indexes. To precisely extract features of different attributes from the generated feature planes we propose a novel attribute indexing method as well as an orthogonal projection regularization to enhance the disentanglement. We also introduce a hyper-latent training strategy and an attribute-specific sampling strategy to avoid style entanglement and misleading punishment from the discriminator. Our method allows users to interactively edit selected attributes in the generated 3D human avatars while keeping others fixed. Both qualitative and quantitative experiments demonstrate that our model provides a strong disentanglement between different attributes allows fine-grained image editing and generates high-quality 3D human avatars.
****
IQ-VFI: Implicit Quadratic Motion Estimation for Video Frame Interpolation
Mengshun Hu, Kui Jiang, Zhihang Zhong, Zheng Wang, Yinqiang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6410-6419
Advanced video frame interpolation (VFI) algorithms approximate intermediate motions between two input frames to synthesize intermediate frame. However they struggle to handle complex scenarios with curvilinear motions since they overlook the latent acceleration information between the input frames. Moreover the supervision of predicted motions is tricky because ground-truth motions are not available. To this end we propose a novel framework for implicit quadratic video frame interpolation (IQ-VFI) which explores latent acceleration information and accurate intermediate motions via knowledge distillation. Specifically the proposed IQ-VFI consists of an implicit acceleration estimation network (IANet) and a VFI backbone the former fully leverages spatio-temporal information to explore latent acceleration priors between two input frames which is then used to progressively modulate linear motions from the latter into quadratic motions in coarse-to-fine manner. Furthermore to encourage both components to distill more acceleration and motion cues oriented towards VFI we propose a knowledge distillation strategy in which implicit acceleration distillation loss and implicit motion distillation loss are employed to adaptively guide latent acceleration priors and intermediate motions learning respectively. Extensive experiments show that our proposed IQ-VFI can achieve state-of-the-art performances on various benchmark datasets.
****

KeyPoint Relative Position Encoding for Face Recognition
Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 244-255
In this paper we address the challenge of making ViT models more robust to unseen affine transformations. Such robustness becomes useful in various recognition tasks such as face recognition when image alignment failures occur. We propose a novel method called KP-RPE which leverages key points (e.g.facial landmarks) to make ViT more resilient to scale translation and pose variations. We begin with the observation that Relative Position Encoding (RPE) is a good way to bring affine transform generalization to ViTs. RPE however can only inject the model with prior knowledge that nearby pixels are more important than far pixels. Keypoint RPE (KP-RPE) is an extension of this principle where the significance of pixels is not solely dictated by their proximity but also by their relative positions to specific keypoints within the image. By anchoring the significance of pixels around keypoints the model can more effectively retain spatial relationships even when those relationships are disrupted by affine transformations. We show the merit of KP-RPE in face and gait recognition. The experimental results demonstrate the effectiveness in improving face recognition performance from low-quality images particularly where alignment is prone to failure. Code and pre-trained models are available.
****
Hyper-MD: Mesh Denoising with Customized Parameters Aware of Noise Intensity and Geometric Characteristics
Xingtao Wang, Hongliang Wei, Xiaopeng Fan, Debin Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4651-4660
Mesh denoising (MD) is a critical task in geometry processing as meshes from scanning or AIGC techniques are susceptible to noise contamination. The challenge of MD lies in the diverse nature of mesh facets in terms of geometric characteristics and noise distributions. Despite recent advancements in deep learning-based MD methods existing MD networks typically neglect the consideration of geometric characteristics and noise distributions. In this paper we propose Hyper-MD a hyper-network-based approach that addresses this limitation by dynamically customizing denoising parameters for each facet based on its noise intensity and geometric characteristics. Specifically Hyper-MD is composed of a hyper-network and an MD network. For each noisy facet the hyper-network takes two angles as input to customize parameters for the MD network. These two angles are specially defined to reveal the noise intensity and geometric characteristics of the current facet respectively. The MD network receives a facet patch as input and outputs the denoised normal using the customized parameters. Experimental results on synthetic and real-scanned meshes demonstrate that Hyper-MD outperforms state-of-the-art mesh denoising methods.
****
Learning Object State Changes in Videos: An Open-World Perspective
Zihui Xue, Kumar Ashutosh, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18493-18503
Object State Changes (OSCs) are pivotal for video understanding. While humans can effortlessly generalize OSC understanding from familiar to unknown objects current approaches are confined to a closed vocabulary. Addressing this gap we introduce a novel open-world formulation for the video OSC problem. The goal is to temporally localize the three stages of an OSC---the object's initial state its transitioning state and its end state---whether or not the object has been observed during training. Towards this end we develop VidOSC a holistic learning approach that: (1) leverages text and vision-language models for supervisory signals to obviate manually labeling OSC training data and (2) abstracts fine-grained shared state representations from objects to enhance generalization. Furthermore we present HowToChange the first open-world benchmark for video OSC localization which offers an order of magnitude increase in the label space and annotation volume compared to the best existing benchmark. Experimental results demonstrate t

he efficacy of our approach in both traditional closed-world and open-world scenarios.
****
Beyond First-Order Tweedie: Solving Inverse Problems using Latent Diffusion
Litu Rout, Yujia Chen, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, Wen-Sheng Chu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9472-9481
Sampling from the posterior distribution in latent diffusion models for inverse problems is computationally challenging. Existing methods often rely on Tweedie's first-order moments that tend to induce biased results. Second-order approximations are computationally prohibitive making standard reverse diffusion processes intractable for posterior sampling. This paper presents Second-order Tweedie sampler from Surrogate Loss (STSL) a novel sampler offering efficiency comparable to first-order Tweedie while enabling tractable reverse processes using second-order approximation. Theoretical results reveal that our approach utilizing for the trace of the Hessian with only O(1) compute establishes a lower bound through a surrogate loss and enables a tractable reverse process. We show STSL outperforms SoTA solvers PSLD and P2L by reducing neural function evaluations by 4X and 8X respectively while enhancing sampling quality on FFHQ ImageNet and COCO benchmarks. Moreover STSL extends to text guided image editing and mitigates residual distortions in corrupted images. To our best knowledge this is the first work to offer an efficient second order approximation for solving inverse problems using latent diffusion and editing real world images with corruptions.
****
Rethinking the Objectives of Vector-Quantized Tokenizers for Image Synthesis
Yuchao Gu, Xintao Wang, Yixiao Ge, Ying Shan, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7631-7640
Vector-Quantized (VQ-based) generative models usually consist of two basic components i.e. VQ tokenizers and generative transformers. Prior research focuses on improving the reconstruction fidelity of VQ tokenizers but rarely examines how the improvement in reconstruction affects the generation ability of generative transformers. In this paper we find that improving the reconstruction fidelity of VQ tokenizers does not necessarily improve the generation. Instead learning to compress semantic features within VQ tokenizers significantly improves generative transformers' ability to capture textures and structures. We thus highlight two competing objectives of VQ tokenizers for image synthesis: semantic compression and details preservation. Different from previous work that prioritizes better details preservation we propose Semantic-Quantized GAN (SeQ-GAN) with two learning phases to balance the two objectives. In the first phase we propose a semantic-enhanced perceptual loss for better semantic compression. In the second phase we fix the encoder and codebook but finetune the decoder to achieve better details preservation. Our proposed SeQ-GAN significantly improves VQ-based generative models for both unconditional and conditional image generation. Specifically SeQ-GAN achieves a Frechet Inception Distance (FID) of 6.25 and Inception Score (IS) of 140.9 on 256x256 ImageNet generation a remarkable improvement over VIT-VQGAN which obtains 11.2 FID and 97.2 IS.
****
ShapeWalk: Compositional Shape Editing Through Language-Guided Chains
Habib Slim, Mohamed Elhoseiny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22574-22583
Editing 3D shapes through natural language instructions is a challenging task that requires the comprehension of both language semantics and fine-grained geometric details. To bridge this gap we introduce ShapeWalk a carefully designed synthetic dataset designed to advance the field of language-guided shape editing. The dataset consists of 158K unique shapes connected through 26K edit chains with an average length of 14 chained shapes. Each consecutive pair of shapes is associated with precise language instructions describing the applied edits. We synthesize edit chains by reconstructing and interpolating shapes sampled from a realistic CAD-designed 3D dataset in the parameter space of the GeoCode shape program

. We leverage rule-based methods and language models to generate accurate and re
alistic natural language prompts corresponding to each edit. To illustrate the p
racticality of our contribution we train neural editor modules in the latent spa
ce of shape autoencoders and demonstrate the ability of our dataset to enable a
variety of language-guided shape edits. Finally we introduce multi-step editing
metrics to benchmark the capacity of our models to perform recursive shape edits
. We hope that our work will enable further study of compositional language-guid
ed shape editing and finds application in 3D CAD design and interactive modeling
.
****
MESA: Matching Everything by Segmenting Anything
Yesheng Zhang, Xu Zhao; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2024, pp. 20217-20226
Feature matching is a crucial task in the field of computer vision which involve
s finding correspondences between images. Previous studies achieve remarkable pe
rformance using learning-based feature comparison. However the pervasive presenc
e of matching redundancy between images gives rise to unnecessary and error-pron
e computations in these methods imposing limitations on their accuracy. To addre
ss this issue we propose MESA a novel approach to establish precise area (or reg
ion) matches for efficient matching redundancy reduction. MESA first leverages t
he advanced image understanding capability of SAM a state-of-the-art foundation
model for image segmentation to obtain image areas with implicit semantic. Then
a multi-relational graph is proposed to model the spatial structure of these are
as and construct their scale hierarchy. Based on graphical models derived from t
he graph the area matching is reformulated as an energy minimization task and ef
fectively resolved. Extensive experiments demonstrate that MESA yields substanti
al precision improvement for multiple point matchers in indoor and outdoor downs
tream tasks e.g. +13.61% for DKM in indoor pose estimation.
****
Learning Degradation-Independent Representations for Camera ISP Pipelines
Yanhui Guo, Fangzhou Luo, Xiaolin Wu; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25774-25783
Image signal processing (ISP) pipeline plays a fundamental role in digital camer
as which converts raw Bayer sensor data to RGB images. However ISP-generated ima
ges usually suffer from imperfections due to the compounded degradations that st
em from sensor noises demosaicing noises compression artifacts and possibly adve
rse effects of erroneous ISP hyperparameter settings such as ISO and gamma value
s. In a general sense these ISP imperfections can be considered as degradations.
The highly complex mechanisms of ISP degradations some of which are even unknow
n pose great challenges to the generalization capability of deep neural networks
(DNN) for image restoration and to their adaptability to downstream tasks. To t
ackle the issues we propose a novel DNN approach to learn degradation-independen
t representations (DiR) through the refinement of a self-supervised learned base
line representation. The proposed DiR learning technique has remarkable domain g
eneralization capability and consequently it outperforms state-of-the-art method
s across various downstream tasks including blind image restoration object detec
tion and instance segmentation as verified in our experiments.
****
SCoFT: Self-Contrastive Fine-Tuning for Equitable Image Generation
Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik
Yun, Andrew Hundt, Jihie Kim, Jean Oh; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10822-10832
Accurate representation in media is known to improve the well-being of the peopl
e who consume it. Generative image models trained on large web-crawled datasets
such as LAION are known to produce images with harmful stereotypes and misrepres
entations of cultures. We improve inclusive representation in generated images b
y (1) engaging with communities to collect a culturally representative dataset t
hat we call the Cross-Cultural Understanding Benchmark (CCUB) and (2) proposing
a novel Self-Contrastive Fine-Tuning (SCoFT pronounced /soft/) method that lever
ages the model's known biases to self-improve. SCoFT is designed to prevent over

fitting on small datasets encode only high-level information from the data and shift the generated distribution away from misrepresentations encoded in a pretrained model. Our user study conducted on 51 participants from 5 different countries based on their self-selected national cultural affiliation shows that fine-tuning on CCUB consistently generates images with higher cultural relevance and fewer stereotypes when compared to the Stable Diffusion baseline which is further improved with our SCoFT technique.
****
Continuous Pose for Monocular Cameras in Neural Implicit Representation
Qi Ma, Danda Pani Paudel, Ajad Chhatkuli, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5291-5301
In this paper we showcase the effectiveness of optimizing monocular camera poses as a continuous function of time. The camera poses are represented using an implicit neural function which maps the given time to the corresponding camera pose. The mapped camera poses are then used for the downstream tasks where joint camera pose optimization is also required. While doing so the network parameters - that implicitly represent camera poses - are optimized. We exploit the proposed method in four diverse experimental settings namely (1) NeRF from noisy poses; (2) NeRF from asynchronous Events; (3) Visual Simultaneous Localization and Mapping (vSLAM); and (4) vSLAM with IMUs. In all four settings the proposed method performs significantly better than the compared baselines and the state-of-the-art methods. Additionally using the assumption of continuous motion changes in pose may actually live in a manifold that has lower than 6 degrees of freedom (DOF) is also realized. We call this low DOF motion representation as the intrinsic motion and use the approach in vSLAM settings show ing impressive camera tracking performance.
****
OmniGlue: Generalizable Feature Matching with Foundation Model Guidance
Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, André Araujo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19865-19875
The image matching field has been witnessing a continuous emergence of novel learnable feature matching techniques with ever-improving performance on conventional benchmarks. However our investigation shows that despite these gains their potential for real-world applications is restricted by their limited generalization capabilities to novel image domains. In this paper we introduce OmniGlue the first learnable image matcher that is designed with generalization as a core principle. OmniGlue leverages broad knowledge from a vision foundation model to guide the feature matching process boosting generalization to domains not seen at training time. Additionally we propose a novel keypoint position-guided attention mechanism which disentangles spatial and appearance information leading to enhanced matching descriptors. We perform comprehensive experiments on a suite of 6 datasets with varied image domains including scene-level object-centric and aerial images. OmniGlue's novel components lead to relative gains on unseen domains of 20.9% with respect to a directly comparable reference model while also outperforming the recent LightGlue method by 9.5% relatively. Code and model can be found at https://hwjiang1510.github.io/OmniGlue.
****
D^4: Dataset Distillation via Disentangled Diffusion Model
Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, Bowen Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5809-5818
Dataset distillation offers a lightweight synthetic dataset for fast network training with promising test accuracy. To imitate the performance of the original dataset most approaches employ bi-level optimization and the distillation space relies on the matching architecture. Nevertheless these approaches either suffer significant computational costs on large-scale datasets or experience performance decline on cross-architectures. We advocate for designing an economical dataset distillation framework that is independent of the matching architectures.With

empirical observations we argue that constraining the consistency of the real and synthetic image spaces will enhance the cross-architecture generalization. Motivated by this we introduce Dataset Distillation via Disentangled Diffusion Model (D^4M) an efficient framework for dataset distillation. Compared to architecture-dependent methods D^4M employs latent diffusion model to guarantee consistency and incorporates label information into category prototypes. The distilled datasets are versatile eliminating the need for repeated generation of distinct datasets for various architectures. Through comprehensive experiments D^4M demonstrates superior performance and robust generalization surpassing the SOTA methods across most aspects.
****
OmniSDF: Scene Reconstruction using Omnidirectional Signed Distance Functions and Adaptive Binoctrees
Hakyeong Kim, Andreas Meuleman, Hyeonjoong Jang, James Tompkin, Min H. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20227-20236
We present a method to reconstruct indoor and outdoor static scene geometry and appearance from an omnidirectional video moving in a small circular sweep. This setting is challenging because of the small baseline and large depth ranges making it difficult to find ray crossings. To better constrain the optimization we estimate geometry as a signed distance field within a spherical binoctree data structure and use a complementary efficient tree traversal strategy based on a breadth-first search for sampling. Unlike regular grids or trees the shape of this structure well-matches the camera setting creating a better memory-quality trade-off. From an initial depth estimate the binoctree is adaptively subdivided throughout the optimization; previous methods use a fixed depth that leaves the scene undersampled. In comparison with three neural optimization methods and two non-neural methods ours shows decreased geometry error on average especially in a detailed scene while significantly reducing the required number of voxels to represent such details.
****
Generating Content for HDR Deghosting from Frequency View
Tao Hu, Qingsen Yan, Yuankai Qi, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25732-25741
Recovering ghost-free High Dynamic Range (HDR) images from multiple Low Dynamic Range (LDR) images becomes challenging when the LDR images exhibit saturation and significant motion. Recent Diffusion Models (DMs) have been introduced in HDR imaging field demonstrating promising performance particularly in achieving visually perceptible results compared to previous DNN-based methods. However DMs require extensive iterations with large models to estimate entire images resulting in inefficiency that hinders their practical application. To address this challenge we propose the Low-Frequency aware Diffusion (LF-Diff) model for ghost-free HDR imaging. The key idea of LF-Diff is implementing the DMs in a highly compacted latent space and integrating it into a regression-based model to enhance the details of reconstructed images. Specifically as low-frequency information is closely related to human visual perception we propose to utilize DMs to create compact low-frequency priors for the reconstruction process. In addition to take full advantage of the above low-frequency priors the Dynamic HDR Reconstruction Network (DHRNet) is carried out in a regression-based manner to obtain final HDR images. Extensive experiments conducted on synthetic and real-world benchmark datasets demonstrate that our LF-Diff performs favorably against several state-of-the-art methods and is 10x faster than previous DM-based methods.
****
Iterated Learning Improves Compositionality in Large Vision-Language Models
Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, Ranjay Krishna; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13785-13795
A fundamental characteristic common to both human vision and natural language is their compositional nature. Yet despite the performance gains contributed by large vision and language pretraining recent investigations find that most--if not

all--our state-of-the-art vision-language models struggle at compositionality. They are unable to distinguish between images of "a girl in white facing a man in black" and "a girl in black facing a man in white". Moreover prior work suggests that compositionality doesn't arise with scale: larger model sizes or training data don't help. This paper develops a new iterated training algorithm that incentivizes compositionality. We draw on decades of cognitive science research that identifies cultural transmission--the need to teach a new generation--as a necessary inductive prior that incentivizes humans to develop compositional languages. Specifically we reframe vision-language contrastive learning as the Lewis Signaling Game between a vision agent and a language agent and operationalize cultural transmission by iteratively resetting one of the agent's weights during training. After every iteration this training paradigm induces representations that become "easier to learn" a property of compositional languages: e.g. our model trained on CC3M and CC12M improves standard CLIP by 4.7% 4.0% respectfully in the SugarCrepe benchmark.
****

Event Stream-based Visual Object Tracking: A High-Resolution Benchmark Dataset and A Novel Baseline

Xiao Wang, Shiao Wang, Chuanming Tang, Lin Zhu, Bo Jiang, Yonghong Tian, Jin Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19248-19257

Tracking with bio-inspired event cameras has garnered increasing interest in recent years. Existing works either utilize aligned RGB and event data for accurate tracking or directly learn an event-based tracker. The former incurs higher inference costs while the latter may be susceptible to the impact of noisy events or sparse spatial resolution. In this paper we propose a novel hierarchical knowledge distillation framework that can fully utilize multi-modal / multi-view information during training to facilitate knowledge transfer enabling us to achieve high-speed and low-latency visual tracking during testing by using only event signals. Specifically a teacher Transformer-based multi-modal tracking framework is first trained by feeding the RGB frame and event stream simultaneously. Then we design a new hierarchical knowledge distillation strategy which includes pairwise similarity feature representation and response maps-based knowledge distillation to guide the learning of the student Transformer network. In particular since existing event-based tracking datasets are all low-resolution (346 * 260) we propose the first large-scale high-resolution (1280 * 720) dataset named EventVOT. It contains 1141 videos and covers a wide range of categories such as pedestrians vehicles UAVs ping pong etc. Extensive experiments on both low-resolution (FE240hz VisEvent COESOT) and our newly proposed high-resolution EventVOT dataset fully validated the effectiveness of our proposed method. The dataset evaluation toolkit and source code will be released.
****

LiDAR-Net: A Real-scanned 3D Point Cloud Dataset for Indoor Scenes

Yanwen Guo, Yuanqi Li, Dayong Ren, Xiaohong Zhang, Jiawei Li, Liang Pu, Changfeng Ma, Xiaoyu Zhan, Jie Guo, Mingqiang Wei, Yan Zhang, Piaopiao Yu, Shuangyu Yang, Donghao Ji, Huisheng Ye, Hao Sun, Yansong Liu, Yinuo Chen, Jiaqi Zhu, Hongyu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21989-21999

In this paper we present LiDAR-Net a new real-scanned indoor point cloud dataset containing nearly 3.6 billion precisely point-level annotated points covering an expansive area of 30000m^2. It encompasses three prevalent daily environments including learning scenes working scenes and living scenes. LiDAR-Net is characterized by its non-uniform point distribution e.g. scanning holes and scanning lines. Additionally it meticulously records and annotates scanning anomalies including reflection noise and ghost. These anomalies stem from specular reflections on glass or metal as well as distortions due to moving persons. LiDAR-Net's realistic representation of non-uniform distribution and anomalies significantly enhances the training of deep learning models leading to improved generalization in practical applications. We thoroughly evaluate the performance of state-of-the-art algorithms on LiDAR-Net and provide a detailed analysis of the results. Cruc

ially our research identifies several fundamental challenges in understanding in door point clouds contributing essential insights to future explorations in this field. Our dataset can be found online: http://lidar-net.njumeta.com
****
Dual DETRs for Multi-Label Temporal Action Detection
Yuhan Zhu, Guozhen Zhang, Jing Tan, Gangshan Wu, Limin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18559-18569
Temporal Action Detection (TAD) aims to identify the action boundaries and the corresponding category within untrimmed videos. Inspired by the success of DETR in object detection several methods have adapted the query-based framework to the TAD task. However these approaches primarily followed DETR to predict actions at the instance level (i.e. identify each action by its center point) leading to sub-optimal boundary localization. To address this issue we propose a new Dual-level query-based TAD framework namely DualDETR to detect actions from both instance-level and boundary-level. Decoding at different levels requires semantics of different granularity therefore we introduce a two-branch decoding structure. This structure builds distinctive decoding processes for different levels facilitating explicit capture of temporal cues and semantics at each level. On top of the two-branch design we present a joint query initialization strategy to align queries from both levels. Specifically we leverage encoder proposals to match queries from each level in a one-to-one manner. Then the matched queries are initialized using position and content prior from the matched action proposal. The aligned dual-level queries can refine the matched proposal with complementary cues during subsequent decoding. We evaluate DualDETR on three challenging multi-label TAD benchmarks. The experimental results demonstrate the superior performance of DualDETR to the existing state-of-the-art methods achieving a substantial improvement under det-mAP and delivering impressive results under seg-mAP.
****
Rich Human Feedback for Text-to-Image Generation
Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katherine M. Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, Vidhya Navalpakkam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19401-19411
Recent Text-to-Image (T2I) generation models such as Stable Diffusion and Imagen have made significant progress in generating high-resolution images based on text descriptions. However many generated images still suffer from issues such as artifacts/implausibility misalignment with text descriptions and low aesthetic quality. Inspired by the success of Reinforcement Learning with Human Feedback (RLHF) for large language models prior works collected human-provided scores as feedback on generated images and trained a reward model to improve the T2I generation. In this paper we enrich the feedback signal by (i) marking image regions that are implausible or misaligned with the text and (ii) annotating which words in the text prompt are misrepresented or missing on the image. We collect such rich human feedback on 18K generated images (RichHF-18K) and train a multimodal transformer to predict the rich feedback automatically. We show that the predicted rich human feedback can be leveraged to improve image generation for example by selecting high-quality training data to finetune and improve the generative models or by creating masks with predicted heatmaps to inpaint the problematic regions. Notably the improvements generalize to models (Muse) beyond those used to generate the images on which human feedback data were collected (Stable Diffusion variants). The RichHF-18K data set will be released in our GitHub repository: https://github.com/google-research/google-research/tree/master/richhf_18k.
****
360DVD: Controllable Panorama Video Generation with 360-Degree Video Diffusion Model
Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6913-6923

Panorama video recently attracts more interest in both study and application cou
rtesy of its immersive experience. Due to the expensive cost of capturing 360-de
gree panoramic videos generating desirable panorama videos by prompts is urgentl
y required. Lately the emerging text-to-video (T2V) diffusion methods demonstrat
e notable effectiveness in standard video generation. However due to the signifi
cant gap in content and motion patterns between panoramic and standard videos th
ese methods encounter challenges in yielding satisfactory 360-degree panoramic v
ideos. In this paper we propose a pipeline named 360-Degree Video Diffusion mode
l (360DVD) for generating 360-degree panoramic videos based on the given prompts
 and motion conditions. Specifically we introduce a lightweight 360-Adapter acco
mpanied by 360 Enhancement Techniques to transform pre-trained T2V models for pa
norama video generation. We further propose a new panorama dataset named WEB360
consisting of panoramic video-text pairs for training 360DVD addressing the abse
nce of captioned panoramic video datasets. Extensive experiments demonstrate the
 superiority and effectiveness of 360DVD for panorama video generation.
****
Map-Relative Pose Regression for Visual Re-Localization
Shuai Chen, Tommaso Cavallari, Victor Adrian Prisacariu, Eric Brachmann; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 20665-20674
Pose regression networks predict the camera pose of a query image relative to a
known environment. Within this family of methods absolute pose regression (APR)
has recently shown promising accuracy in the range of a few centimeters in posit
ion error. APR networks encode the scene geometry implicitly in their weights. T
o achieve high accuracy they require vast amounts of training data that realisti
cally can only be created using novel view synthesis in a days-long process. Thi
s process has to be repeated for each new scene again and again. We present a ne
w approach to pose regression map-relative pose regression (marepo) that satisfi
es the data hunger of the pose regression network in a scene-agnostic fashion. W
e condition the pose regressor on a scene-specific map representation such that
its pose predictions are relative to the scene map. This allows us to train the
pose regressor across hundreds of scenes to learn the generic relation between a
 scene-specific map representation and the camera pose. Our map-relative pose re
gressor can be applied to new map representations immediately or after mere minu
tes of fine-tuning for the highest accuracy. Our approach outperforms previous p
ose regression methods by far on two public datasets indoor and outdoor. Code is
 available: https://nianticlabs.github.io/marepo.
****
Implicit Event-RGBD Neural SLAM
Delin Qu, Chi Yan, Dong Wang, Jie Yin, Qizhi Chen, Dan Xu, Yiting Zhang, Bin Zha
o, Xuelong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2024, pp. 19584-19594
Implicit neural SLAM has achieved remarkable progress recently. Nevertheless exi
sting methods face significant challenges in non-ideal scenarios such as motion
blur or lighting variation which often leads to issues like convergence failures
 localization drifts and distorted mapping. To address these challenges we propo
se EN-SLAM the first event-RGBD implicit neural SLAM framework which effectively
 leverages the high rate and high dynamic range advantages of event data for tra
cking and mapping. Specifically EN-SLAM proposes a differentiable CRF (Camera Re
sponse Function) rendering technique to generate distinct RGB and event camera d
ata via a shared radiance field which is optimized by learning a unified implici
t representation with the captured event and RGBD supervision. Moreover based on
 the temporal difference property of events we propose a temporal aggregating op
timization strategy for the event joint tracking and global bundle adjustment ca
pitalizing on the consecutive difference constraints of events significantly enh
ancing tracking accuracy and robustness. Finally we construct the simulated data
set DEV-Indoors and real captured dataset DEV-Reals containing 6 scenes 17 seque
nces with practical motion blur and lighting changes for evaluations. Experiment
al results show that our method outperforms the SOTA methods in both tracking AT
E and mapping ACC with a real-time 17 FPS in various challenging environments. P

roject page: https://delinqu.github.io/EN-SLAM.
****
Virtual Immunohistochemistry Staining for Histological Images Assisted by Weakly-supervised Learning
Jiahan Li, Jiuyang Dong, Shenjin Huang, Xi Li, Junjun Jiang, Xiaopeng Fan, Yongbing Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11259-11268
Recently virtual staining technology has greatly promoted the advancement of histopathology. Despite the practical successes achieved the outstanding performance of most virtual staining methods relies on hard-to-obtain paired images in training. In this paper we propose a method for virtual immunohistochemistry (IHC) staining named confusion-GAN which does not require paired images and can achieve comparable performance to supervised algorithms. Specifically we propose a multi-branch discriminator which judges if the features of generated images can be embedded into the feature pool of target domain images to improve the visual quality of generated images. Meanwhile we also propose a novel patch-level pathology information extractor which is assisted by multiple instance learning to ensure pathological consistency during virtual staining. Extensive experiments were conducted on three types of IHC images including a high-resolution hepatocellular carcinoma immunohistochemical dataset proposed by us. The results demonstrated that our proposed confusion-GAN can generate highly realistic images that are capable of deceiving even experienced pathologists. Furthermore compared to using H&E images directly the downstream diagnosis achieved higher accuracy when using images generated by confusion-GAN. Our dataset and codes will be available at https://github.com/jiahanli2022/confusion-GAN.
****
DeCoTR: Enhancing Depth Completion with 2D and 3D Attentions
Yunxiao Shi, Manish Kumar Singh, Hong Cai, Fatih Porikli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10736-10746
In this paper we introduce a novel approach that harnesses both 2D and 3D attentions to enable highly accurate depth completion without requiring iterative spatial propagations. Specifically we first enhance a baseline convolutional depth completion model by applying attention to 2D features in the bottleneck and skip connections. This effectively improves the performance of this simple network and sets it on par with the latest complex transformer-based models. Leveraging the initial depths and features from this network we uplift the 2D features to form a 3D point cloud and construct a 3D point transformer to process it allowing the model to explicitly learn and exploit 3D geometric features. In addition we propose normalization techniques to process the point cloud which improves learning and leads to better accuracy than directly using point transformers off the shelf. Furthermore we incorporate global attention on downsampled point cloud features which enables long-range context while still being computationally feasible. We evaluate our method DeCoTR on established depth completion benchmarks including NYU Depth V2 and KITTI showcasing that it sets new state-of-the-art performance. We further conduct zero-shot evaluations on ScanNet and DDAD benchmarks and demonstrate that DeCoTR has superior generalizability compared to existing approaches.
****
Utility-Fairness Trade-Offs and How to Find Them
Sepehr Dehdashtian, Bashir Sadeghi, Vishnu Naresh Boddeti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12037-12046
When building classification systems with demographic fairness considerations there are two objectives to satisfy: 1) maximizing utility for the specific task and 2) ensuring fairness w.r.t. a known demographic attribute. These objectives often compete so optimizing both can lead to a trade-off between utility and fairness. While existing works acknowledge the trade-offs and study their limits two questions remain unanswered: 1) What are the optimal tradeoffs between utility and fairness? and 2) How can we numerically quantify these trade-offs from data

for a desired prediction task and demographic attribute of interest? This paper addresses these questions. We introduce two utility-fairness trade-offs: the Data-Space and Label-Space Trade-off. The trade-offs reveal three regions within the utility-fairness plane delineating what is fully and partially possible and impossible. We propose U-FaTE a method to numerically quantify the trade-offs for a given prediction task and group fairness definition from data samples. Based on the trade-offs we introduce a new scheme for evaluating representations. An extensive evaluation of fair representation learning methods and representations from over 1000 pre-trained models revealed that most current approaches are far from the estimated and achievable fairness-utility trade-offs across multiple datasets and prediction tasks.
****
Domain-Specific Block Selection and Paired-View Pseudo-Labeling for Online Test-Time Adaptation
Yeonguk Yu, Sungho Shin, Seunghyeok Back, Mihwan Ko, Sangjun Noh, Kyoobin Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22723-22732
Test-time adaptation (TTA) aims to adapt a pre-trained model to a new test domain without access to source data after deployment. Existing approaches typically rely on self-training with pseudo-labels since ground-truth cannot be obtained from test data. Although the quality of pseudo labels is important for stable and accurate long-term adaptation it has not been previously addressed. In this work we propose DPLOT a simple yet effective TTA framework that consists of two components: (1) domain-specific block selection and (2) pseudo-label generation using paired-view images. Specifically we select blocks that involve domain-specific feature extraction and train these blocks by entropy minimization. After blocks are adjusted for current test domain we generate pseudo-labels by averaging given test images and corresponding flipped counterparts. By simply using flip augmentation we prevent a decrease in the quality of the pseudo-labels which can be caused by the domain gap resulting from strong augmentation. Our experimental results demonstrate that DPLOT outperforms previous TTA methods in CIFAR10-C CIFAR100-C and ImageNet-C benchmarks reducing error by up to 5.4% 9.1% and 2.9% respectively. Also we provide an extensive analysis to demonstrate effectiveness of our framework. Code is available at https://github.com/gist-ailab/domain-specific-block-selection-and-paired-view-pseudo-labeling-for-online-TTA.
****
Aerial Lifting: Neural Urban Semantic and Building Instance Lifting from Aerial Imagery
Yuqi Zhang, Guanying Chen, Jiaxing Chen, Shuguang Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21092-21103
We present a neural radiance field method for urban-scale semantic and building-level instance segmentation from aerial images by lifting noisy 2D labels to 3D. This is a challenging problem due to two primary reasons. Firstly objects in urban aerial images exhibit substantial variations in size including buildings cars and roads which pose a significant challenge for accurate 2D segmentation. Secondly the 2D labels generated by existing segmentation methods suffer from the multi-view inconsistency problem especially in the case of aerial images where each image captures only a small portion of the entire scene. To overcome these limitations we first introduce a scale-adaptive semantic label fusion strategy that enhances the segmentation of objects of varying sizes by combining labels predicted from different altitudes harnessing the novel-view synthesis capabilities of NeRF. We then introduce a novel cross-view instance label grouping strategy based on the 3D scene representation to mitigate the multi-view inconsistency problem in the 2D instance labels. Furthermore we exploit multi-view reconstructed depth priors to improve the geometric quality of the reconstructed radiance field resulting in enhanced segmentation results. Experiments on multiple real-world urban-scale datasets demonstrate that our approach outperforms existing methods highlighting its effectiveness. The source code is available at https://github.com/zyqz97/Aerial_lifting.

****
SAOR: Single-View Articulated Object Reconstruction
Mehmet Aygun, Oisin Mac Aodha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10382-10391
We introduce SAOR a novel approach for estimating the 3D shape texture and viewpoint of an articulated object from a single image captured in the wild. Unlike prior approaches that rely on pre-defined category-specific 3D templates or tailored 3D skeletons SAOR learns to articulate shapes from single-view image collections with a skeleton-free part-based model without requiring any 3D object shape priors. To prevent ill-posed solutions we propose a cross-instance consistency loss that exploits disentangled object shape deformation and articulation. This is helped by a new silhouette-based sampling mechanism to enhance viewpoint diversity during training. Our method only requires estimated object silhouettes and relative depth maps from off-the-shelf pre-trained networks during training. At inference time given a single-view image it efficiently outputs an explicit mesh representation. We obtain improved qualitative and quantitative results on challenging quadruped animals compared to relevant existing work.
****
A Theory of Joint Light and Heat Transport for Lambertian Scenes
Mani Ramanagopal, Sriram Narayanan, Aswin C. Sankaranarayanan, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11924-11933
We present a novel theory that establishes the relationship between light transport in visible and thermal infrared and heat transport in solids. We show that heat generated due to light absorption can be estimated by modeling heat transport using a thermal camera. For situations where heat conduction is negligible we analytically solve the heat transport equation to derive a simple expression relating the change in thermal image intensity to the absorbed light intensity and heat capacity of the material. Next we prove that intrinsic image decomposition for Lambertian scenes becomes a well-posed problem if one has access to the absorbed light. Our theory generalizes to arbitrary shapes and unstructured illumination. Our theory is based on applying energy conservation principle at each pixel independently. We validate our theory using real-world experiments on diffuse objects made of different materials that exhibit both direct and global components (inter-reflections) of light transport under unknown complex lighting.
****
iKUN: Speak to Trackers without Retraining
Yunhao Du, Cheng Lei, Zhicheng Zhao, Fei Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19135-19144
Referring multi-object tracking (RMOT) aims to track multiple objects based on input textual descriptions. Previous works realize it by simply integrating an extra textual module into the multi-object tracker. However they typically need to retrain the entire framework and have difficulties in optimization. In this work we propose an insertable Knowledge Unification Network termed iKUN to enable communication with off-the-shelf trackers in a plug-and-play manner. Concretely a knowledge unification module (KUM) is designed to adaptively extract visual features based on textual guidance. Meanwhile to improve the localization accuracy we present a neural version of Kalman filter (NKF) to dynamically adjust process noise and observation noise based on the current motion status. Moreover to address the problem of open-set long-tail distribution of textual descriptions a test-time similarity calibration method is proposed to refine the confidence score with pseudo frequency. Extensive experiments on Refer-KITTI dataset verify the effectiveness of our framework. Finally to speed up the development of RMOT we also contribute a more challenging dataset Refer-Dance by extending public DanceTrack dataset with motion and dressing descriptions. The codes and dataset are available at https://github.com/dyhBUPT/iKUN.
****
RankMatch: Exploring the Better Consistency Regularization for Semi-supervised Semantic Segmentation
Huayu Mai, Rui Sun, Tianzhu Zhang, Feng Wu; Proceedings of the IEEE/CVF Conferen

ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3391-3401
The key lie in semi-supervised semantic segmentation is how to fully exploit substantial unlabeled data to improve the model's generalization performance by resorting to constructing effective supervision signals. Most methods tend to directly apply contrastive learning to seek additional supervision to complement independent regular pixel-wise consistency regularization. However these methods tend not to be preferred ascribed to their complicated designs heavy memory footprints and susceptibility to confirmation bias. In this paper we analyze the bottlenecks exist in contrastive learning-based methods and offer a fresh perspective on inter-pixel correlations to construct more safe and effective supervision signals which is in line with the nature of semantic segmentation. To this end we develop a coherent RankMatch network including the construction of representative agents to model inter-pixel correlation beyond regular individual pixel-wise consistency and further unlock the potential of agents by modeling inter-agent relationships in pursuit of rank-aware correlation consistency. Extensive experimental results on multiple benchmarks including mitochondria segmentation demonstrate that RankMatch performs favorably against state-of-the-art methods. Particularly in the low-data regimes RankMatch achieves significant improvements.
****
Facial Identity Anonymization via Intrinsic and Extrinsic Attention Distraction
Zhenzhong Kuang, Xiaochen Yang, Yingjie Shen, Chao Hu, Jun Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12406-12415
The unprecedented capture and application of face images raise increasing concerns on anonymization to fight against privacy disclosure. Most existing methods may suffer from the problem of excessive change of the identity-independent information or insufficient identity protection. In this paper we present a new face anonymization approach by distracting the intrinsic and extrinsic identity attentions. On the one hand we anonymize the identity information in the feature space by distracting the intrinsic identity attention. On the other we anonymize the visual clues (i.e. appearance and geometry structure) by distracting the extrinsic identity attention. Our approach allows for flexible and intuitive manipulation of face appearance and geometry structure to produce diverse results and it can also be used to instruct users to perform personalized anonymization. We conduct extensive experiments on multiple datasets and demonstrate that our approach outperforms state-of-the-art methods.
****
3D-SceneDreamer: Text-Driven 3D-Consistent Scene Generation
Songchun Zhang, Yibo Zhang, Quan Zheng, Rui Ma, Wei Hua, Hujun Bao, Weiwei Xu, Changqing Zou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10170-10180
Text-driven 3D scene generation techniques have made rapid progress in recent years. Their success is mainly attributed to using existing generative models to iteratively perform image warping and inpainting to generate 3D scenes. However these methods heavily rely on the outputs of existing models leading to error accumulation in geometry and appearance that prevent the models from being used in various scenarios (e.g. outdoor and unreal scenarios). To address this limitation we generatively refine the newly generated local views by querying and aggregating global 3D information and then progressively generate the 3D scene. Specifically we employ a tri-plane features-based NeRF as a unified representation of the 3D scene to constrain global 3D consistency and propose a generative refinement network to synthesize new contents with higher quality by exploiting the natural image prior from 2D diffusion model as well as the global 3D information of the current scene. Our extensive experiments demonstrate that in comparison to previous methods our approach supports wide variety of scene generation and arbitrary camera trajectories with improved visual quality and 3D consistency.
****
VMINer: Versatile Multi-view Inverse Rendering with Near- and Far-field Light Sources
Fan Fei, Jiajun Tang, Ping Tan, Boxin Shi; Proceedings of the IEEE/CVF Conferenc

e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11800-11809
This paper introduces a versatile multi-view inverse rendering framework with near- and far-field light sources. Tackling the fundamental challenge of inherent ambiguity in inverse rendering our framework adopts a lightweight yet inclusive lighting model for different near- and far-field lights thus is able to make use of input images under varied lighting conditions available during capture. It leverages observations under each lighting to disentangle the intrinsic geometry and material from the external lighting using both neural radiance field rendering and physically-based surface rendering on the 3D implicit fields. After training the reconstructed scene is extracted to a textured triangle mesh for seamless integration into industrial rendering software for various applications. Quantitatively and qualitatively tested on synthetic and real-world scenes our method shows superiority to state-of-the-art multi-view inverse rendering methods in both speed and quality.
****
RoHM: Robust Human Motion Reconstruction via Diffusion
Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, Federica Bogo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14606-14617
We propose RoHM an approach for robust 3D human motion reconstruction from monocular RGB(-D) videos in the presence of noise and occlusions. Most previous approaches either train neural networks to directly regress motion in 3D or learn data-driven motion priors and combine them with optimization at test time. RoHM is a novel diffusion-based motion model that conditioned on noisy and occluded input data reconstructs complete plausible motions in consistent global coordinates. Given the complexity of the problem -- requiring one to address different tasks (denoising and infilling) in different solution spaces (local and global motion) -- we decompose it into two sub-tasks and learn two models one for global trajectory and one for local motion. To capture the correlations between the two we then introduce a novel conditioning module combining it with an iterative inference scheme. We apply RoHM to a variety of tasks -- from motion reconstruction and denoising to spatial and temporal infilling. Extensive experiments on three popular datasets show that our method outperforms state-of-the-art approaches qualitatively and quantitatively while being faster at test time. The code is available at https://sanweiliti.github.io/ROHM/ROHM.html.
****
Do You Remember? Dense Video Captioning with Cross-Modal Memory Retrieval
Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, Seong Tae Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13894-13904
There has been significant attention to the research on dense video captioning which aims to automatically localize and caption all events within untrimmed video. Several studies introduce methods by designing dense video captioning as a multitasking problem of event localization and event captioning to consider inter-task relations. However addressing both tasks using only visual input is challenging due to the lack of semantic content. In this study we address this by proposing a novel framework inspired by the cognitive information processing of humans. Our model utilizes external memory to incorporate prior knowledge. The memory retrieval method is proposed with cross-modal video-to-text matching. To effectively incorporate retrieved text features the versatile encoder and the decoder with visual and textual cross-attention modules are designed. Comparative experiments have been conducted to show the effectiveness of the proposed method on ActivityNet Captions and YouCook2 datasets. Experimental results show promising performance of our model without extensive pretraining from a large video dataset. Our code is available at https://github.com/ailab-kyunghee/CM2_DVC.
****
DuPL: Dual Student with Trustworthy Progressive Learning for Robust Weakly Supervised Semantic Segmentation
Yuanchen Wu, Xichen Ye, Kequan Yang, Jide Li, Xiaoqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.

3534-3543
Recently One-stage Weakly Supervised Semantic Segmentation (WSSS) with image-lev
el labels has gained increasing interest due to simplification over its cumberso
me multi-stage counterpart. Limited by the inherent ambiguity of Class Activatio
n Map (CAM) we observe that one-stage pipelines often encounter confirmation bia
s caused by incorrect CAM pseudo-labels impairing their final segmentation perfo
rmance. Although recent works discard many unreliable pseudo-labels to implicitl
y alleviate this issue they fail to exploit sufficient supervision for their mod
els. To this end we propose a dual student framework with trustworthy progressiv
e learning (DuPL). Specifically we propose a dual student network with a discrep
ancy loss to yield diverse CAMs for each sub-net. The two sub-nets generate supe
rvision for each other mitigating the confirmation bias caused by learning their
 own incorrect pseudo-labels. In this process we progressively introduce more tr
ustworthy pseudo-labels to be involved in the supervision through dynamic thresh
old adjustment with an adaptive noise filtering strategy. Moreover we believe th
at every pixel even discarded from supervision due to its unreliability is impor
tant for WSSS. Thus we develop consistency regularization on these discarded reg
ions providing supervision of every pixel. Experiment results demonstrate the su
periority of the proposed DuPL over the recent state-of-the-art alternatives on
PASCAL VOC 2012 and MS COCO datasets. Code is available at https://github.com/Wu
0409/DuPL.
****
Learning with Structural Labels for Learning with Noisy Labels
Noo-ri Kim, Jin-Seop Lee, Jee-Hyong Lee; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27610-27620
Deep Neural Networks (DNNs) have demonstrated remarkable performance across dive
rse domains and tasks with large-scale datasets. To reduce labeling costs for la
rge-scale datasets semi-automated and crowdsourcing labeling methods are develop
ed but their labels are inevitably noisy. Learning with Noisy Labels (LNL) appro
aches aim to train DNNs despite the presence of noisy labels. These approaches u
tilize the memorization effect to select correct labels and refine noisy ones wh
ich are then used for subsequent training. However these methods encounter a sig
nificant decrease in the model's generalization performance due to the inevitabl
y existing noise labels. To overcome this limitation we propose a new approach t
o enhance learning with noisy labels by incorporating additional distribution in
formation--structural labels. In order to leverage additional distribution infor
mation for generalization we employ a reverse k-NN which helps the model in achi
eving a better feature manifold and mitigating overfitting to noisy labels. The
proposed method shows outperformed performance in multiple benchmark datasets wi
th IDN and real-world noisy datasets.
****
SurMo: Surface-based 4D Motion Modeling for Dynamic Human Rendering
Tao Hu, Fangzhou Hong, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2024, pp. 6550-6560
Dynamic human rendering from video sequences has achieved remarkable progress by
 formulating the rendering as a mapping from static poses to human images. Howev
er existing methods focus on the human appearance reconstruction of every single
 frame while the temporal motion relations are not fully explored. In this paper
 we propose a new 4D motion modeling paradigm SurMo that jointly models the temp
oral dynamics and human appearances in a unified framework with three key design
s: 1) Surface-based motion encoding that models 4D human motions with an efficie
nt compact surface-based triplane. It encodes both spatial and temporal motion r
elations on the dense surface manifold of a statistical body template which inhe
rits body topology priors for generalizable novel view synthesis with sparse tra
ining observations. 2) Physical motion decoding that is designed to encourage ph
ysical motion learning by decoding the motion triplane features at timestep t to
 predict both spatial derivatives and temporal derivatives at the next timestep
t+1 in the training stage. 3) 4D appearance decoding that renders the motion tri
planes into images by an efficient volumetric surface-conditioned renderer that
focuses on the rendering of body surfaces with motion learning conditioning. Ext

ensive experiments validate the state-of-the-art performance of our new paradigm and illustrate the expressiveness of surface-based motion triplanes for rendering high-fidelity view-consistent humans with fast motions and even motion-dependent shadows. Our project page is at: https://taohuumd.github.io/projects/SurMo.
****
SPAD: Spatially Aware Multi-View Diffusers
Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10026-10038
We present SPAD a novel approach for creating consistent multi-view images from text prompts or single images. To enable multi-view generation we repurpose a pretrained 2D diffusion model by extending its self-attention layers with cross-view interactions and fine-tune it on a high quality subset of Objaverse. We find that a naive extension of the self-attention proposed in prior work (e.g. MVDream) leads to content copying between views. Therefore we explicitly constrain the cross-view attention based on epipolar geometry. To further enhance 3D consistency we utilize Pl ?ucker coordinates derived from camera rays and inject them as positional encoding. This enables SPAD to reason over spatial proximity in 3D well. Compared to concurrent works that can only generate views at fixed azimuth and elevation (e.g. MVDream SyncDreamer) SPAD offers full camera control and achieves state-of-the-art results in novel view synthesis on unseen objects from the Objaverse and Google Scanned Objects datasets. Finally we demonstrate that text-to-3D generation using SPAD prevents the multi-face Janus issue.
****
Gradient Reweighting: Towards Imbalanced Class-Incremental Learning
Jiangpeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16668-16677
Class-Incremental Learning (CIL) trains a model to continually recognize new classes from non-stationary data while retaining learned knowledge. A major challenge of CIL arises when applying to real-world data characterized by non-uniform distribution which introduces a dual imbalance problem involving (i) disparities between stored exemplars of old tasks and new class data (inter-phase imbalance) and (ii) severe class imbalances within each individual task (intra-phase imbalance). We show that this dual imbalance issue causes skewed gradient updates with biased weights in FC layers thus inducing over/under-fitting and catastrophic forgetting in CIL. Our method addresses it by reweighting the gradients towards balanced optimization and unbiased classifier learning. Additionally we observe imbalanced forgetting where paradoxically the instance-rich classes suffer higher performance degradation during CIL due to a larger amount of training data becoming unavailable in subsequent learning phases. To tackle this we further introduce a distribution-aware knowledge distillation loss to mitigate forgetting by aligning output logits proportionally with the distribution of lost training data. We validate our method on CIFAR-100 ImageNetSubset and Food101 across various evaluation protocols and demonstrate consistent improvements compared to existing works showing great potential to apply CIL in real-world scenarios with enhanced robustness and effectiveness.
****
Hierarchical Spatio-temporal Decoupling for Text-to-Video Generation
Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, Nong Sang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6635-6645
Despite diffusion models having shown powerful abilities to generate photorealistic images generating videos that are realistic and diverse still remains in its infancy. One of the key reasons is that current methods intertwine spatial content and temporal dynamics together leading to a notably increased complexity of text-to-video generation (T2V). In this work we propose HiGen a diffusion model-based method that improves performance by decoupling the spatial and temporal factors of videos from two perspectives i.e. structure level and content level. At the structure level we decompose the T2V task into two steps including spatial

reasoning and temporal reasoning using a unified denoiser. Specifically we generate spatially coherent priors using text during spatial reasoning and then generate temporally coherent motions from these priors during temporal reasoning. At the content level we extract two subtle cues from the content of the input video that can express motion and appearance changes respectively. These two cues then guide the model's training for generating videos enabling flexible content variations and enhancing temporal stability. Through the decoupled paradigm HiGen can effectively reduce the complexity of this task and generate realistic videos with semantics accuracy and motion stability. Extensive experiments demonstrate the superior performance of HiGen over the state-of-the-art T2V methods. We have released our source code and models.
****
PLACE: Adaptive Layout-Semantic Fusion for Semantic Image Synthesis
Zhengyao Lv, Yuxiang Wei, Wangmeng Zuo, Kwan-Yee K. Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9264-9274
Recent advancements in large-scale pre-trained text-to-image models have led to remarkable progress in semantic image synthesis. Nevertheless synthesizing high-quality images with consistent semantics and layout remains a challenge. In this paper we propose the adaPtive LAyout-semantiC fusion modulE (PLACE) that harnesses pre-trained models to alleviate the aforementioned issues. Specifically we first employ the layout control map to faithfully represent layouts in the feature space. Subsequently we combine the layout and semantic features in a timestep-adaptive manner to synthesize images with realistic details. During fine-tuning we propose the Semantic Alignment (SA) loss to further enhance layout alignment. Additionally we introduce the Layout-Free Prior Preservation (LFP) loss which leverages unlabeled data to maintain the priors of pre-trained models thereby improving the visual quality and semantic consistency of synthesized images. Extensive experiments demonstrate that our approach performs favorably in terms of visual quality semantic consistency and layout alignment. The source code and model are available at \href https://github.com/cszy98/PLACE/tree/main PLACE .
****
Exploring Efficient Asymmetric Blind-Spots for Self-Supervised Denoising in Real-World Scenarios
Shiyan Chen, Jiyuan Zhang, Zhaofei Yu, Tiejun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2814-2823
Self-supervised denoising has attracted widespread attention due to its ability to train without clean images. However noise in real-world scenarios is often spatially correlated which causes many self-supervised algorithms that assume pixel-wise independent noise to perform poorly. Recent works have attempted to break noise correlation with downsampling or neighborhood masking. However denoising on downsampled subgraphs can lead to aliasing effects and loss of details due to a lower sampling rate. Furthermore the neighborhood masking methods either come with high computational complexity or do not consider local spatial preservation during inference. Through the analysis of existing methods we point out that the key to obtaining high-quality and texture-rich results in real-world self-supervised denoising tasks is to train at the original input resolution structure and use asymmetric operations during training and inference. Based on this we propose Asymmetric Tunable Blind-Spot Network (AT-BSN) where the blind-spot size can be freely adjusted thus better balancing noise correlation suppression and image local spatial destruction during training and inference. In addition we regard the pre-trained AT-BSN as a meta-teacher network capable of generating various teacher networks by sampling different blind-spots. We propose a blind-spot based multi-teacher distillation strategy to distill a lightweight network significantly improving performance. Experimental results on multiple datasets prove that our method achieves state-of-the-art and is superior to other self-supervised algorithms in terms of computational overhead and visual effects.
****
Gaussian Splatting SLAM

Hidenobu Matsuki, Riku Murai, Paul H.J. Kelly, Andrew J. Davison; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 18039-18048
We present the first application of 3D Gaussian Splatting in monocular SLAM the
most fundamental but the hardest setup for Visual SLAM. Our method which runs li
ve at 3fps utilises Gaussians as the only 3D representation unifying the require
d representation for accurate efficient tracking mapping and high-quality render
ing. Designed for challenging monocular settings our approach is seamlessly exte
ndable to RGB-D SLAM when an external depth sensor is available. Several innovat
ions are required to continuously reconstruct 3D scenes with high fidelity from
a live camera. First to move beyond the original 3DGS algorithm which requires a
ccurate poses from an offline Structure from Motion (SfM) system we formulate ca
mera tracking for 3DGS using direct optimisation against the 3D Gaussians and sh
ow that this enables fast and robust tracking with a wide basin of convergence.
Second by utilising the explicit nature of the Gaussians we introduce geometric
verification and regularisation to handle the ambiguities occurring in increment
al 3D dense reconstruction. Finally we introduce a full SLAM system which not on
ly achieves state-of-the-art results in novel view synthesis and trajectory esti
mation but also reconstruction of tiny and even transparent objects.
****
Not All Classes Stand on Same Embeddings: Calibrating a Semantic Distance with M
etric Tensor
Jae Hyeon Park, Gyoomin Lee, Seunggi Park, Sung In Cho; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1772
2-17731
The consistency training (CT)-based semi-supervised learning (SSL) bites state-o
f-the-art performance on SSL-based image classification. However the existing CT
-based SSL methods do not highlight the non-Euclidean characteristics and class-
wise varieties of embedding spaces in an SSL model thus they cannot fully utiliz
e the effectiveness of CT. Thus we propose a metric tensor-based consistency reg
ularization exploiting the class-variant geometrical structure of embeddings on
the high-dimensional feature space. The proposed method not only minimizes the p
rediction discrepancy between different views of a given image but also estimate
s the intrinsic geometric curvature of embedding spaces by employing the global
and local metric tensors. The global metric tensor is used to globally estimate
the class-invariant embeddings from the whole data distribution while the local
metric tensor is exploited to estimate the class-variant embeddings of each clus
ter. The two metric tensors are optimized by the consistency regularization base
d on the weak and strong augmentation strategy. The proposed method provides the
 highest classification accuracy on average compared to the existing state-of-th
e-art SSL methods on conventional datasets.
****
A Simple Recipe for Contrastively Pre-training Video-First Encoders Beyond 16 Fr
ames
Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, V
iorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, Aida Nematzdeh;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2024, pp. 14386-14397
Understanding long real-world videos requires modeling of long-range visual depe
ndencies. To this end we explore video-first architectures building on the commo
n paradigm of transferring large-scale image--text models to video via shallow t
emporal fusion. However we expose two limitations to the approach: (1) decreased
 spatial capabilities likely due to poor video--language alignment in standard v
ideo datasets and (2) higher memory consumption bottlenecking the number of fram
es that can be processed. To mitigate the memory bottleneck we systematically an
alyze the memory/accuracy trade-off of various efficient methods: factorized att
ention parameter-efficient image-to-video adaptation input masking and multi-res
olution patchification. Surprisingly simply masking large portions of the video
(up to 75%) during contrastive pre-training proves to be one of the most robust
ways to scale encoders to videos up to 4.3 minutes at 1 FPS. Our simple approach

for training long video-to-text models which scales to 1B parameters does not add new architectural complexity and is able to outperform the popular paradigm of using much larger LLMs as an information aggregator over segment-based information on benchmarks with long-range temporal dependencies (YouCook2 EgoSchema).
****
DeMatch: Deep Decomposition of Motion Field for Two-View Correspondence Learning
Shihua Zhang, Zizhuo Li, Yuan Gao, Jiayi Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20278-20287
Two-view correspondence learning has recently focused on considering the coherence and smoothness of the motion field between an image pair. Dominant schemes include controlling the complexity of the field function with regularization or smoothing the field with local filters but the former suffers from heavy computational burden and the latter fails to accommodate discontinuities in the case of large scene disparities. In this paper inspired by Fourier expansion we propose a novel network called DeMatch which decomposes the motion field to retain its main "low-frequency" and smooth part. This achieves implicit regularization with lower computational cost and generates piecewise smoothness naturally. Specifically we first decompose the rough motion field that is contaminated by false matches into several different sub-fields which are highly smooth and contain the main energy of the original field. Then with these smooth sub-fields we recover a cleaner motion field from which correct motion vectors are subsequently derived. We also design a special masked decomposition strategy to further mitigate the negative influence of false matches. All the mentioned processes are finally implemented in a discrete and learnable manner avoiding the difficulty of calculating real dense fields. Extensive experiments reveal that DeMatch outperforms state-of-the-art methods in multiple tasks and shows promising low computational usage and piecewise smoothness property. The code and trained models are publicly available at https://github.com/SuhZhang/DeMatch.
****
Hierarchical Diffusion Policy for Kinematics-Aware Multi-Task Robotic Manipulation
Xiao Ma, Sumit Patidar, Iain Haughton, Stephen James; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18081-18090
This paper introduces Hierarchical Diffusion Policy (HDP) a hierarchical agent for multi-task robotic manipulation. HDP factorises a manipulation policy into a hierarchical structure: a high-level task-planning agent which predicts a distant next-best end-effector pose (NBP) and a low-level goal-conditioned diffusion policy which generates optimal motion trajectories. The factorised policy representation allows HDP to tackle both long-horizon task planning while generating fine-grained low-level actions. To generate context-aware motion trajectories while satisfying robot kinematics constraints we present a novel kinematics-aware goal-conditioned control agent Robot Kinematics Diffuser (RK-Diffuser). Specifically RK-Diffuser learns to generate both the end-effector pose and joint position trajectories and distill the accurate but kinematics-unaware end-effector pose diffuser to the kinematics-aware but less accurate joint position diffuser via differentiable kinematics. Empirically we show that HDP achieves a significantly higher success rate than the state-of-the-art methods in both simulation and real-world.
****
Efficient Multi-scale Network with Learnable Discrete Wavelet Transform for Blind Motion Deblurring
Xin Gao, Tianheng Qiu, Xinyu Zhang, Hanlin Bai, Kang Liu, Xuan Huang, Hu Wei, Guoying Zhang, Huaping Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2733-2742
Coarse-to-fine schemes are widely used in traditional single-image motion deblur; however in the context of deep learning existing multi-scale algorithms not only require the use of complex modules for feature fusion of low-scale RGB images and deep semantics but also manually generate low-resolution pairs of images that do not have sufficient confidence. In this work we propose a multi-scale netw

ork based on single-input and multiple-outputs(SIMO) for motion deblurring. This simplifies the complexity of algorithms based on a coarse-to-fine scheme. To alleviate restoration defects impacting detail information brought about by using a multi-scale architecture we combine the characteristics of real-world blurring trajectories with a learnable wavelet transform module to focus on the directional continuity and frequency features of the step-by-step transitions between blurred images to sharp images. In conclusion we propose a multi-scale network with a learnable discrete wavelet transform (MLWNet) which exhibits state-of-the-art performance on multiple real-world deblurred datasets in terms of both subjective and objective quality as well as computational efficiency.
****
MaskPLAN: Masked Generative Layout Planning from Partial Input
Hang Zhang, Anton Savov, Benjamin Dillenburger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8964-8973
Layout planning spanning from architecture to interior design is a slow iterative exploration of ill-defined problems adopting a "I'll know it when I see it" approach to potential solutions. Recent advances in generative models promise automating layout generation yet often overlook the crucial role of user-guided iteration cannot generate full solutions from incomplete design ideas and do not learn for the inter-dependency of layout attributes. To address these limitations we propose MaskPLAN a novel generative model based on Graph-structured Dynamic Masked Autoencoders (GDMAE) featuring five transformers generating a blend of graph-based and image-based layout attributes. MaskPLAN lets users generate and adjust layouts with partial attribute definitions create alternatives for preferences and practice new composition-driven or functionality-driven workflows. Through cross-attribute learning and the user input as a global conditional prior we ensure that design synthesis is calibrated at every intermediate stage maintaining its feasibility and practicality. Extensive evaluations show MaskPLAN's superior performance over existing methods across multiple metrics.
****
Benchmarking the Robustness of Temporal Action Detection Models Against Temporal Corruptions
Runhao Zeng, Xiaoyong Chen, Jiaming Liang, Huisi Wu, Guangzhong Cao, Yong Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18263-18274
Temporal action detection (TAD) aims to locate action positions and recognize action categories in long-term untrimmed videos. Although many methods have achieved promising results their robustness has not been thoroughly studied. In practice we observe that temporal information in videos can be occasionally corrupted such as missing or blurred frames. Interestingly existing methods often incur a significant performance drop even if only one frame is affected. To formally evaluate the robustness we establish two temporal corruption robustness benchmarks namely THUMOS14-C and ActivityNet-v1.3-C. In this paper we extensively analyze the robustness of seven leading TAD methods and obtain some interesting findings: 1) Existing methods are particularly vulnerable to temporal corruptions and end-to-end methods are often more susceptible than those with a pre-trained feature extractor; 2) Vulnerability mainly comes from localization error rather than classification error; 3) When corruptions occur in the middle of an action instance TAD models tend to yield the largest performance drop. Besides building a benchmark we further develop a simple but effective robust training method to defend against temporal corruptions through the FrameDrop augmentation and Temporal-Robust Consistency loss. Remarkably our approach not only improves robustness but also yields promising improvements on clean data. We believe that this study will serve as a benchmark for future research in robust video analysis. Source code and models are available at https://github.com/Alvin-Zeng/temporal-robustness-benchmark.
****
Open-World Human-Object Interaction Detection via Multi-modal Prompts
Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, Ruimao Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp

. 16954-16964
In this paper we develop MP-HOI a powerful Multi-modal Prompt-based HOI detector designed to leverage both textual descriptions for open-set generalization and visual exemplars for handling high ambiguity in descriptions realizing HOI detection in the open world. Specifically it integrates visual prompts into existing language-guided-only HOI detectors to handle situations where textual descriptions face difficulties in generalization and to address complex scenarios with high interaction ambiguity. To facilitate MP-HOI training we build a large-scale HOI dataset named Magic-HOI which gathers six existing datasets into a unified label space forming over 186K images with 2.4K objects 1.2K actions and 20K HOI interactions. Furthermore to tackle the long-tail issue within the Magic-HOI dataset we introduce an automated pipeline for generating realistically annotated HOI images and present SynHOI a high-quality synthetic HOI dataset containing 100K images. Leveraging these two datasets MP-HOI optimizes the HOI task as a similarity learning process between multi-modal prompts and objects/interactions via a unified contrastive loss to learn generalizable and transferable objects/interactions representations from large-scale data. MP-HOI could serve as a generalist HOI detector surpassing the HOI vocabulary of existing expert models by more than 30 times. Concurrently our results demonstrate that MP-HOI exhibits remarkable zero-shot capability in real-world scenarios and consistently achieves a new state-of-the-art performance across various benchmarks. Our project homepage is available at https://MP-HOI.github.io/.
****
HMD-Poser: On-Device Real-time Human Motion Tracking from Scalable Sparse Observations
Peng Dai, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, Zeming Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 874-884
It is especially challenging to achieve real-time human motion tracking on a standalone VR Head-Mounted Display (HMD) such as Meta Quest and PICO. In this paper we propose HMD-Poser the first unified approach to recover full-body motions using scalable sparse observations from HMD and body-worn IMUs. In particular it can support a variety of input scenarios such as HMD HMD+2IMUs HMD+3IMUs etc. The scalability of inputs may accommodate users' choices for both high tracking accuracy and easy-to-wear. A lightweight temporal-spatial feature learning network is proposed in HMD-Poser to guarantee that the model runs in real-time on HMDs. Furthermore HMD-Poser presents online body shape estimation to improve the position accuracy of body joints. Extensive experimental results on the challenging AMASS dataset show that HMD-Poser achieves new state-of-the-art results in both accuracy and real-time performance. We also build a new free-dancing motion dataset to evaluate HMD-Poser's on-device performance and investigate the performance gap between synthetic data and real-captured sensor data. Finally we demonstrate our HMD-Poser with a real-time Avatar-driving application on a commercial HMD. Our code and free-dancing motion dataset are available \href https://pico-ai-team.github.io/hmd-poser here .
****
UniMODE: Unified Monocular 3D Object Detection
Zhuoling Li, Xiaogang Xu, SerNam Lim, Hengshuang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16561-16570
Realizing unified monocular 3D object detection including both indoor and outdoor scenes holds great importance in applications like robot navigation. However involving various scenarios of data to train models poses challenges due to their significantly different characteristics e.g. diverse geometry properties and heterogeneous domain distributions. To address these challenges we build a detector based on the bird's-eye-view (BEV) detection paradigm where the explicit feature projection is beneficial to addressing the geometry learning ambiguity when employing multiple scenarios of data to train detectors. Then we split the classical BEV detection architecture into two stages and propose an uneven BEV grid design to handle the convergence instability caused by the aforementioned challeng

es. Moreover we develop a sparse BEV feature projection strategy to reduce computational cost and a unified domain alignment method to handle heterogeneous domains. Combining these techniques a unified detector UniMODE is derived which surpasses the previous state-of-the-art on the challenging Omni3D dataset (a large-scale dataset including both indoor and outdoor scenes) by 4.9% \rm AP_ 3D revealing the first successful generalization of a BEV detector to unified 3D object detection.
****
Sherpa3D: Boosting High-Fidelity Text-to-3D Generation via Coarse 3D Prior
Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, Yueqi Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20763-20774
Recently 3D content creation from text prompts has demonstrated remarkable progress by utilizing 2D and 3D diffusion models. While 3D diffusion models ensure great multi-view consistency their ability to generate high-quality and diverse 3D assets is hindered by the limited 3D data. In contrast 2D diffusion models find a distillation approach that achieves excellent generalization and rich details without any 3D data. However 2D lifting methods suffer from inherent view-agnostic ambiguity thereby leading to serious multi-face Janus issues where text prompts fail to provide sufficient guidance to learn coherent 3D results. Instead of retraining a costly viewpoint-aware model we study how to fully exploit easily accessible coarse 3D knowledge to enhance the prompts and guide 2D lifting optimization for refinement. In this paper we propose Sherpa3D a new text-to-3D framework that achieves high-fidelity generalizability and geometric consistency simultaneously. Specifically we design a pair of guiding strategies derived from the coarse 3D prior generated by the 3D diffusion model: a structural guidance for geometric fidelity and a semantic guidance for 3D coherence. Employing the two types of guidance the 2D diffusion model enriches the 3D content with diversified and high-quality results. Extensive experiments show the superiority of our Sherpa3D over the state-of-the-art text-to-3D methods in terms of quality and 3D consistency.
****
Flexible Biometrics Recognition: Bridging the Multimodality Gap through Attention Alignment and Prompt Tuning
Leslie Ching Ow Tiong, Dick Sigmund, Chen-Hui Chan, Andrew Beng Jin Teoh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 267-276
Periocular and face are complementary biometrics for identity management albeit with inherent limitations notably in scenarios involving occlusion due to sunglasses or masks. In response to these challenges we introduce Flexible Biometric Recognition (FBR) a novel framework designed to advance conventional face periocular and multimodal face-periocular biometrics across both intra- and cross-modality recognition tasks. FBR strategically utilizes the Multimodal Fusion Attention (MFA) and Multimodal Prompt Tuning (MPT) mechanisms within the Vision Transformer architecture. MFA facilitates the fusion of modalities ensuring cohesive alignment between facial and periocular embeddings while incorporating soft-biometrics to enhance the model's ability to discriminate between individuals. The fusion of three modalities is pivotal in exploring interrelationships between different modalities. Additionally MPT serves as a unifying bridge intertwining inputs and promoting cross-modality interactions while preserving their distinctive characteristics. The collaborative synergy of MFA and MPT enhances the shared features of the face and periocular with a specific emphasis on the ocular region yielding exceptional performance in both intra- and cross-modality recognition tasks. Rigorous experimentation across four benchmark datasets validates the noteworthy performance of the FBR model. The source code is available at https://github.com/MIS-DevWorks/FBR.
****
Multi-agent Collaborative Perception via Motion-aware Robust Communication Network
Shixin Hong, Yu Liu, Zhi Li, Shaohui Li, You He; Proceedings of the IEEE/CVF Con

ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15301-15310
Collaborative perception allows for information sharing between multiple agents such as vehicles and infrastructure to obtain a comprehensive view of the enviro nment through communication and fusion. Current research on multi-agent collabor ative perception systems often assumes ideal communication and perception enviro nments and neglects the effect of real-world noise such as pose noise motion blu r and perception noise. To address this gap in this paper we propose a novel mot ion-aware robust communication network (MRCNet) that mitigates noise interferenc e and achieves accurate and robust collaborative perception. MRCNet consists of two main components: multi-scale robust fusion (MRF) addresses pose noise by dev eloping cross-semantic multi-scale enhanced aggregation to fuse features of diff erent scales while motion enhanced mechanism (MEM) captures motion context to co mpensate for information blurring caused by moving objects. Experimental results on popular collaborative 3D object detection datasets demonstrate that MRCNet o utperforms competing methods in noisy scenarios with improved perception perform ance using less bandwidth.
****
The Manga Whisperer: Automatically Generating Transcriptions for Comics
Ragav Sachdeva, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2024, pp. 12967-12976
In the past few decades Japanese comics commonly referred to as Manga have trans cended both cultural and linguistic boundaries to become a true worldwide sensat ion. Yet the inherent reliance on visual cues and illustration within manga rend ers it largely inaccessible to individuals with visual impairments. In this work we seek to address this substantial barrier with the aim of ensuring that manga can be appreciated and actively engaged by everyone. Specifically we tackle the problem of diarisation i.e. generating a transcription of who said what and whe n in a fully automatic way. To this end we make the following contributions: (1) we present a unified model Magi that is able to (a) detect panels text boxes an d character boxes (b) cluster characters by identity (without knowing the number of clusters apriori) and (c) associate dialogues to their speakers; (2) we prop ose a novel approach that is able to sort the detected text boxes in their readi ng order and generate a dialogue transcript; (3) we annotate an evaluation bench mark for this task using publicly available [English] manga pages.
****
Exploring Region-Word Alignment in Built-in Detector for Open-Vocabulary Object Detection
Heng Zhang, Qiuyu Zhao, Linyu Zheng, Hao Zeng, Zhiwei Ge, Tianhao Li, Sulong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2024, pp. 16975-16984
Open-vocabulary object detection aims to detect novel categories that are indepe ndent from the base categories used during training. Most modern methods adhere to the paradigm of learning vision-language space from a large-scale multi-modal corpus and subsequently transferring the acquired knowledge to off-the-shelf de tectors like Faster-RCNN. However information attenuation or destruction may occ ur during the process of knowledge transfer due to the domain gap hampering the generalization ability on novel categories. To mitigate this predicament in this paper we present a novel framework named BIND standing for Bulit-IN Detector to eliminate the need for module replacement or knowledge transfer to off-the-shel f detectors. Specifically we design a two-stage training framework with an Encod er-Decoder structure. In the first stage an image-text dual encoder is trained t o learn region-word alignment from a corpus of image-text pairs. In the second s tage a DETR-style decoder is trained to perform detection on annotated object de tection datasets. In contrast to conventional manually designed non-adaptive anc hors which generate numerous redundant proposals we develop an anchor proposal n etwork that generates anchor proposals with high likelihood based on candidates adaptively thereby substantially improving detection efficiency. Experimental re sults on two public benchmarks COCO and LVIS demonstrate that our method stands as a state-of-the-art approach for open-vocabulary object detection.
****

MovieChat: From Dense Token to Sparse Memory for Long Video Understanding
Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, Gaoang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18221-18232
Recently integrating video foundation models and large language models to build a video understanding system can overcome the limitations of specific pre-defined vision tasks. Yet existing systems can only handle videos with very few frames. For long videos the computation complexity memory cost and long-term temporal connection impose additional challenges. Taking advantage of the Atkinson-Shiffrin memory model with tokens in Transformers being employed as the carriers of memory in combination with our specially designed memory mechanism we propose the MovieChat to overcome these challenges. MovieChat achieves state-of-the-art performance in long video understanding along with the released MovieChat-1K benchmark with 1K long video and 14K manual annotations for validation of the effectiveness of our method. The code models and data can be found in https://rese1f.github.io/MovieChat.
****
Comparing the Decision-Making Mechanisms by Transformers and CNNs via Explanation Methods
Mingqi Jiang, Saeed Khorram, Li Fuxin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9546-9555
In order to gain insights about the decision-making of different visual recognition backbones we propose two methodologies sub-explanation counting and cross-testing that systematically applies deep explanation algorithms on a dataset-wide basis and compares the statistics generated from the amount and nature of the explanations. These methodologies reveal the difference among networks in terms of two properties called compositionality and disjunctivism. Transformers and ConvNeXt are found to be more compositional in the sense that they jointly consider multiple parts of the image in building their decisions whereas traditional CNNs and distilled transformers are less compositional and more disjunctive which means that they use multiple diverse but smaller set of parts to achieve a confident prediction. Through further experiments we pinpointed the choice of normalization to be especially important in the compositionality of a model in that batch normalization leads to less compositionality while group and layer normalization lead to more. Finally we also analyze the features shared by different backbones and plot a landscape of different models based on their feature-use similarity.
****
A Unified Diffusion Framework for Scene-aware Human Motion Estimation from Sparse Signals
Jiangnan Tang, Jingya Wang, Kaiyang Ji, Lan Xu, Jingyi Yu, Ye Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21251-21262
Estimating full-body human motion via sparse tracking signals from head-mounted displays and hand controllers in 3D scenes is crucial to applications in AR/VR. One of the biggest challenges to this task is the one-to-many mapping from sparse observations to dense full-body motions which endowed inherent ambiguities. To help resolve this ambiguous problem we introduce a new framework to combine rich contextual information provided by scenes to benefit full-body motion tracking from sparse observations. To estimate plausible human motions given sparse tracking signals and 3D scenes we develop \text S ^2Fusion a unified framework fusing \underline S cene and sparse \underline S ignals with a conditional dif\underline Fusion model. \text S ^2Fusion first extracts the spatial-temporal relations residing in the sparse signals via a periodic autoencoder and then produces time-alignment feature embedding as additional inputs. Subsequently by drawing initial noisy motion from a pre-trained prior \text S ^2Fusion utilizes conditional diffusion to fuse scene geometry and sparse tracking signals to generate full-body scene-aware motions. The sampling procedure of \text S ^2Fusion is further guided by a specially designed scene-penetration loss and phase-matching loss whi

ch effectively regularizes the motion of the lower body even in the absence of a
ny tracking signals making the generated motion much more plausible and coherent
. Extensive experimental results have demonstrated that our \text S ^2Fusion out
performs the state-of-the-art in terms of estimation quality and smoothness.
****
Single Domain Generalization for Crowd Counting
Zhuoxuan Peng, S.-H. Gary Chan; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2024, pp. 28025-28034
Due to its promising results density map regression has been widely employed for
 image-based crowd counting. The approach however often suffers from severe perf
ormance degradation when tested on data from unseen scenarios the so-called "dom
ain shift" problem. To address the problem we investigate in this work single do
main generalization (SDG) for crowd counting. The existing SDG approaches are ma
inly for image classification and segmentation and can hardly be extended to our
 case due to its regression nature and label ambiguity (i.e. ambiguous pixel-lev
el ground truths). We propose MPCount a novel effective SDG approach even for na
rrow source distribution. MPCount stores diverse density values for density map
regression and reconstructs domain-invariant features by means of only one memor
y bank a content error mask and attention consistency loss. By partitioning the
image into grids it employs patch-wise classification as an auxiliary task to mi
tigate label ambiguity. Through extensive experiments on different datasets MPCo
unt is shown to significantly improve counting accuracy compared to the state of
 the art under diverse scenarios unobserved in the training data characterized b
y narrow source distribution. Code is available at https://github.com/Shimmer93/
MPCount.
****
Atlantis: Enabling Underwater Depth Estimation with Stable Diffusion
Fan Zhang, Shaodi You, Yu Li, Ying Fu; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11852-11861
Monocular depth estimation has experienced significant progress on terrestrial i
mages in recent years thanks to deep learning advancements. But it remains inade
quate for underwater scenes primarily due to data scarcity. Given the inherent c
hallenges of light attenuation and backscatter in water acquiring clear underwat
er images or precise depth is notably difficult and costly. To mitigate this iss
ue learning-based approaches often rely on synthetic data or turn to self- or un
supervised manners. Nonetheless their performance is often hindered by domain ga
p and looser constraints. In this paper we propose a novel pipeline for generati
ng photorealistic underwater images using accurate terrestrial depth. This appro
ach facilitates the supervised training of models for underwater depth estimatio
n effectively reducing the performance disparity between terrestrial and underwa
ter environments. Contrary to previous synthetic datasets that merely apply styl
e transfer to terrestrial images without scene content change our approach uniqu
ely creates vivid non-existent underwater scenes by leveraging terrestrial depth
 data through the innovative Stable Diffusion model. Specifically we introduce a
 specialized Depth2Underwater ControlNet trained on prepared \ Underwater Depth
Text\ data triplets for this generation task. Our newly developed dataset Atlan
tis enables terrestrial depth estimation models to achieve considerable improvem
ents on unseen underwater scenes surpassing their terrestrial pretrained counter
parts both quantitatively and qualitatively. Moreover we further show its practi
cal utility by applying the improved depth in underwater image enhancement and i
ts smaller domain gap from the LLVM perspective. Code and dataset are publicly a
vailable at https://github.com/zkawfanx/Atlantis.
****
Matching Anything by Segmenting Anything
Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool
, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2024, pp. 18963-18973
The robust association of the same objects across video frames in complex scenes
 is crucial for many applications especially object tracking. Current methods pr
edominantly rely on labeled domain-specific video datasets which limits cross-do

main generalization of learned similarity embeddings. We propose MASA a novel method for robust instance association learning capable of matching any objects within videos across diverse domains without tracking labels. Leveraging the rich object segmentation from the Segment Anything Model (SAM) MASA learns instance-level correspondence through exhausive data transformations. We treat the SAM outputs as dense object region proposals and learn to match those regions from a vast image collection. We further design a universal MASA adapter which can work in tandem with foundational segmentation or detection models and enable them to track any detected objects. Those combinations present strong zero-shot tracking ability in complex domains. Extensive tests on multiple challenging MOT and MOTS benchmarks indicate that the proposed method using only unlabelled static images achieves even better performance than state-of-the-art methods trained with fully annotated in-domain video sequences in zero-shot association. Our code is available at https://github.com/siyuanliii/masa.
****
Task-Aware Encoder Control for Deep Video Compression
Xingtong Ge, Jixiang Luo, Xinjie Zhang, Tongda Xu, Guo Lu, Dailan He, Jing Geng, Yan Wang, Jun Zhang, Hongwei Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26036-26045
Prior research on deep video compression (DVC) for machine tasks typically necessitates training a unique codec for each specific task mandating a dedicated decoder per task. In contrast traditional video codecs employ a flexible encoder controller enabling the adaptation of a single codec to different tasks through mechanisms like mode prediction. Drawing inspiration from this we introduce an innovative encoder controller for deep video compression for machines. This controller features a mode prediction and a Group of Pictures (GoP) selection module. Our approach centralizes control at the encoding stage allowing for adaptable encoder adjustments across different tasks such as detection and tracking while maintaining compatibility with a standard pre-trained DVC decoder. Empirical evidence demonstrates that our method is applicable across multiple tasks with various existing pre-trained DVCs. Moreover extensive experiments demonstrate that our method outperforms previous DVC by about 25% bitrate for different tasks with only one pre-trained decoder.
****
Multi-scale Dynamic and Hierarchical Relationship Modeling for Facial Action Units Recognition
Zihan Wang, Siyang Song, Cheng Luo, Songhe Deng, Weicheng Xie, Linlin Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1270-1280
Human facial action units (AUs) are mutually related in a hierarchical manner as not only they are associated with each other in both spatial and temporal domains but also AUs located in the same/close facial regions show stronger relationships than those of different facial regions. While none of existing approach thoroughly model such hierarchical inter-dependencies among AUs this paper proposes to comprehensively model multi-scale AU-related dynamic and hierarchical spatio-temporal relationship among AUs for their occurrences recognition. Specifically we first propose a novel multi-scale temporal differencing network with an adaptive weighting block to explicitly capture facial dynamics across frames at different spatial scales which specifically considers the heterogeneity of range and magnitude in different AUs' activation. Then a two-stage strategy is introduced to hierarchically model the relationship among AUs based on their spatial distribution (i.e. local and cross-region AU relationship modelling). Experimental results achieved on BP4D and DISFA show that our approach is the new state-of-the-art in the field of AU occurrence recognition. Our code is publicly available at https://github.com/CVI-SZU/MDHR.
****
Decoupled Pseudo-labeling for Semi-Supervised Monocular 3D Object Detection
Jiacheng Zhang, Jiaming Li, Xiangru Lin, Wei Zhang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, Guanbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16923-16932

We delve into pseudo-labeling for semi-supervised monocular 3D object detection (SSM3OD) and discover two primary issues: a misalignment between the prediction quality of 3D and 2D attributes and the tendency of depth supervision derived fr om pseudo-labels to be noisy leading to significant optimization conflicts with other reliable forms of supervision. To tackle these issues we introduce a novel decoupled pseudo-labeling (DPL) approach for SSM3OD. Our approach features a De coupled Pseudo-label Generation (DPG) module designed to efficiently generate ps eudo-labels by separately processing 2D and 3D attributes. This module incorpora tes a unique homography-based method for identifying dependable pseudo-labels in Bird's Eye View (BEV) space specifically for 3D attributes. Additionally we pre sent a Depth Gradient Projection (DGP) module to mitigate optimization conflicts caused by noisy depth supervision of pseudo-labels effectively decoupling the d epth gradient and removing conflicting gradients. This dual decoupling strategy--at both the pseudo-label generation and gradient levels--significantly improves the utilization of pseudo-labels in SSM3OD. Our comprehensive experiments on th e KITTI benchmark demonstrate the superiority of our method over existing approa ches.
****
Temporally Consistent Unbalanced Optimal Transport for Unsupervised Action Segme ntation
Ming Xu, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2024, pp. 14618-14627
We propose a novel approach to the action segmentation task for long untrimmed v ideos based on solving an optimal transport problem. By encoding a temporal cons istency prior into a Gromov-Wasserstein problem we are able to decode a temporal ly consistent segmentation from a noisy affinity/matching cost matrix between vi deo frames and action classes. Unlike previous approaches our method does not re quire knowing the action order for a video to attain temporal consistency. Furth ermore our resulting (fused) Gromov-Wasserstein problem can be efficiently solve d on GPUs using a few iterations of projected mirror descent. We demonstrate the effectiveness of our method in an unsupervised learning setting where our metho d is used to generate pseudo-labels for self-training. We evaluate our segmentat ion approach and unsupervised learning pipeline on the Breakfast 50-Salads YouTu be Instructions and Desktop Assembly datasets yielding state-of-the-art results for the unsupervised video action segmentation task.
****
Learning Transferable Negative Prompts for Out-of-Distribution Detection
Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, Jin Zheng; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17584-17594
Existing prompt learning methods have shown certain capabilities in Out-of-Distr ibution (OOD) detection but the lack of OOD images in the target dataset in thei r training can lead to mismatches between OOD images and In-Distribution (ID) ca tegories resulting in a high false positive rate. To address this issue we intro duce a novel OOD detection method named 'NegPrompt' to learn a set of negative p rompts each representing a negative connotation of a given class label for delin eating the boundaries between ID and OOD images. It learns such negative prompts with ID data only without any reliance on external outlier data. Further curren t methods assume the availability of samples of all ID classes rendering them in effective in open-vocabulary learning scenarios where the inference stage can co ntain novel ID classes not present during training. In contrast our learned nega tive prompts are transferable to novel class labels. Experiments on various Imag eNet benchmarks show that NegPrompt surpasses state-of-the-art prompt-learning-b ased OOD detection methods and maintains a consistent lead in hard OOD detection in closed- and open-vocabulary classification scenarios. Code is available at h ttps://github.com/mala-lab/negprompt.
****
Long-Tail Class Incremental Learning via Independent Sub-prototype Construction
Xi Wang, Xu Yang, Jie Yin, Kun Wei, Cheng Deng; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28598-28607

Long-tail class incremental learning (LT-CIL) is designed to perpetually acquire novel knowledge from an imbalanced and perpetually evolving data stream while ensuring the retention of previously acquired knowledge. The existing method only re-balances data distribution and ignores exploring the potential relationship between different samples causing non-robust representations and even severe forgetting in classes with few samples. In this paper we constructed two parallel spaces simultaneously: 1) Sub-prototype space and 2) Reminiscence space to learn robust representations while alleviating forgetfulness. Concretely we advance the concept of the sub-prototype space which amalgamates insights from diverse classes. This integration facilitates the mutual complementarity of varied knowledge thereby augmenting the attainment of more robust representations. Furthermore we introduce the reminiscence space which encapsulates each class distribution aiming to constraint model optimization and mitigate the phenomenon of forgetting. The tandem utilization of the two parallel spaces effectively alleviates the adverse consequences associated with imbalanced data distribution preventing forgetting without needing replay examples. Extensive experiments demonstrate that our method achieves state-of-the-art performance on various benchmarks.
****
Learning with Unreliability: Fast Few-shot Voxel Radiance Fields with Relative Geometric Consistency
Yingjie Xu, Bangzhen Liu, Hao Tang, Bailin Deng, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20342-20351
We propose a voxel-based optimization framework ReVoRF for few-shot radiance fields that strategically addresses the unreliability in pseudo novel view synthesis. Our method pivots on the insight that relative depth relationships within neighboring regions are more reliable than the absolute color values in disoccluded areas. Consequently we devise a bilateral geometric consistency loss that carefully navigates the trade-off between color fidelity and geometric accuracy in the context of depth consistency for uncertain regions. Moreover we present a reliability-guided learning strategy to discern and utilize the variable quality across synthesized views complemented by a reliability-aware voxel smoothing algorithm that smoothens the transition between reliable and unreliable data patches. Our approach allows for a more nuanced use of all available data promoting enhanced learning from regions previously considered unsuitable for high-quality reconstruction. Extensive experiments across diverse datasets reveal that our approach attains significant gains in efficiency and accuracy delivering rendering speeds of 3 FPS 7 mins to train a 360deg scene and a 5% improvement in PSNR over existing few-shot methods. Code is available at https://github.com/HKCLynn/ReVoRF
****
Towards Understanding and Improving Adversarial Robustness of Vision Transformers
Samyak Jain, Tanima Dutta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24736-24745
Recent literature has demonstrated that vision transformers (VITs) exhibit superior performance compared to convolutional neural networks (CNNs). The majority of recent research on adversarial robustness however has predominantly focused on CNNs. In this work we bridge this gap by analyzing the effectiveness of existing attacks on VITs. We demonstrate that due to the softmax computations in every attention block in VITs they are inherently vulnerable to floating point underflow errors. This can lead to a gradient masking effect resulting in suboptimal attack strength of well-known attacks like PGD Carlini and Wagner (CW) GAMA and Patch attacks. Motivated by this we propose Adaptive Attention Scaling (AAS) attack that can automatically find the optimal scaling factors of pre-softmax outputs using gradient-based optimization. We show that the proposed simple strategy can be incorporated with any existing adversarial attacks as well as adversarial training methods and achieved improved performance. On VIT-B16 we demonstrate an improved attack strength of upto 2.2% on CIFAR10 and upto 2.9% on CIFAR100 by incorporating the proposed AAS attack with state-of-the-art single attack methods like GAMA attack. Further we utilise the proposed AAS attack for every few epoch

s in existing adversarial training methods which is termed as Adaptive Attention Scaling Adversarial Training (AAS-AT). On incorporating AAS-AT with existing methods we outperform them on VITs over 1.3-3.5% on CIFAR10. We observe improved performance on ImageNet-100 as well.
****
EventEgo3D: 3D Human Motion Capture from Egocentric Event Streams
Christen Millerdurai, Hiroyasu Akada, Jian Wang, Diogo Luvizon, Christian Theobalt, Vladislav Golyanik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1186-1195
Monocular egocentric 3D human motion capture is a challenging and actively researched problem. Existing methods use synchronously operating visual sensors (e.g. RGB cameras) and often fail under low lighting and fast motions which can be restricting in many applications involving head-mounted devices. In response to the existing limitations this paper 1) introduces a new problem i.e. 3D human motion capture from an egocentric monocular event camera with a fisheye lens and 2) proposes the first approach to it called EventEgo3D (EE3D). Event streams have high temporal resolution and provide reliable cues for 3D human motion capture under high-speed human motions and rapidly changing illumination. The proposed EE3D framework is specifically tailored for learning with event streams in the LNES representation enabling high 3D reconstruction accuracy. We also design a prototype of a mobile head-mounted device with an event camera and record a real dataset with event observations and the ground-truth 3D human poses (in addition to the synthetic dataset). Our EE3D demonstrates robustness and superior 3D accuracy compared to existing solutions across various challenging experiments while supporting real-time 3D pose update rates of 140Hz.
****
Holistic Features are almost Sufficient for Text-to-Video Retrieval
Kaibin Tian, Ruixiang Zhao, Zijie Xin, Bangxiang Lan, Xirong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17138-17147
For text-to-video retrieval (T2VR) which aims to retrieve unlabeled videos by ad-hoc textual queries CLIP-based methods currently lead the way. Compared to CLIP4Clip which is efficient and compact state-of-the-art models tend to compute video-text similarity through fine-grained cross-modal feature interaction and matching putting their scalability for large-scale T2VR applications into doubt. We propose TeachCLIP enabling a CLIP4Clip based student network to learn from more advanced yet computationally intensive models. In order to create a learning channel to convey fine-grained cross-modal knowledge from a heavy model to the student we add to CLIP4Clip a simple Attentional frame-Feature Aggregation (AFA) block which by design adds no extra storage / computation overhead at the retrieval stage. Frame-text relevance scores calculated by the teacher network are used as soft labels to supervise the attentive weights produced by AFA. Extensive experiments on multiple public datasets justify the viability of the proposed method. TeachCLIP has the same efficiency and compactness as CLIP4Clip yet has near-SOTA effectiveness.
****
A Call to Reflect on Evaluation Practices for Age Estimation: Comparative Analysis of the State-of-the-Art and a Unified Benchmark
Jakub Paplhám, Vojt?ch Franc; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1196-1205
Comparing different age estimation methods poses a challenge due to the unreliability of published results stemming from inconsistencies in the benchmarking process. Previous studies have reported continuous performance improvements over the past decade using specialized methods; however our findings challenge these claims. This paper identifies two trivial yet persistent issues with the currently used evaluation protocol and describes how to resolve them. We offer an extensive comparative analysis for state-of-the-art facial age estimation methods. Surprisingly we find that the performance differences between the methods are negligible compared to the effect of other factors such as facial alignment facial coverage image resolution model architecture or the amount of data used for pretrai

ning. We use the gained insights to propose using FaRL as the backbone model and demonstrate its effectiveness on all public datasets. We make the source code and exact data splits public on GitHub and in the supplementary material.
****
CosalPure: Learning Concept from Group Images for Robust Co-Saliency Detection
Jiayi Zhu, Qing Guo, Felix Juefei-Xu, Yihao Huang, Yang Liu, Geguang Pu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3669-3678
Co-salient object detection (CoSOD) aims to identify the common and salient (usually in the foreground) regions across a given group of images. Although achieving significant progress state-of-the-art CoSODs could be easily affected by some adversarial perturbations leading to substantial accuracy reduction. The adversarial perturbations can mislead CoSODs but do not change the high-level semantic information (e.g. concept) of the co-salient objects. In this paper we propose a novel robustness enhancement framework by first learning the concept of the co-salient objects based on the input group images and then leveraging this concept to purify adversarial perturbations which are subsequently fed to CoSODs for robustness enhancement. Specifically we propose CosalPure containing two modules i.e. group-image concept learning and concept-guided diffusion purification. For the first module we adopt a pre-trained text-to-image diffusion model to learn the concept of co-salient objects within group images where the learned concept is robust to adversarial examples. For the second module we map the adversarial image to the latent space and then perform diffusion generation by embedding the learned concept into the noise prediction function as an extra condition. Our method can effectively alleviate the influence of the SOTA adversarial attack containing different adversarial patterns including exposure and noise. The extensive results demonstrate that our method could enhance the robustness of CoSODs significantly.
****
Uncertainty-aware Action Decoupling Transformer for Action Anticipation
Hongji Guo, Nakul Agarwal, Shao-Yuan Lo, Kwonjoon Lee, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18644-18654
Human action anticipation aims at predicting what people will do in the future based on past observations. In this paper we introduce Uncertainty-aware Action Decoupling Transformer (UADT) for action anticipation. Unlike existing methods that directly predict action in a verb-noun pair format we decouple the action anticipation task into verb and noun anticipations separately. The objective is to make the two decoupled tasks assist each other and eventually improve the action anticipation task. Specifically we propose a two-stream Transformer-based architecture which is composed of a verb-to-noun model and a noun-to-verb model. The verb-to-noun model leverages the verb information to improve the noun prediction and the other way around. We extend the model in a probabilistic manner and quantify the predictive uncertainty of each decoupled task to select features. In this way the noun prediction leverages the most informative and redundancy-free verb features and verb prediction works similarly. Finally the two streams are combined dynamically based on their uncertainties to make the joint action anticipation. We demonstrate the efficacy of our method by achieving state-of-the-art performance on action anticipation benchmarks including EPIC-KITCHENS EGTEA Gaze+ and 50-Salads.
****
MRFP: Learning Generalizable Semantic Segmentation from Sim-2-Real with Multi-Resolution Feature Perturbation
Sumanth Udupa, Prajwal Gurunath, Aniruddh Sikdar, Suresh Sundaram; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5904-5914
Deep neural networks have shown exemplary performance on semantic scene understanding tasks on source domains but due to the absence of style diversity during training enhancing performance on unseen target domains using only single source domain data remains a challenging task. Generation of simulated data is a feasib

le alternative to retrieving large style-diverse real-world datasets as it is a cumbersome and budget-intensive process. However the large domain-specific inconsistencies between simulated and real-world data pose a significant generalization challenge in semantic segmentation. In this work to alleviate this problem we propose a novel Multi-Resolution Feature Perturbation (MRFP) technique to randomize domain-specific fine-grained features and perturb style of coarse features. Our experimental results on various urban-scene segmentation datasets clearly indicate that along with the perturbation of style-information perturbation of fine-feature components is paramount to learn domain invariant robust feature maps for semantic segmentation models. MRFP is a simple and computationally efficient transferable module with no additional learnable parameters or objective functions that helps state-of-the-art deep neural networks to learn robust domain invariant features for simulation-to-real semantic segmentation. Code is available at https://github.com/airl-iisc/MRFP.
****
S-DyRF: Reference-Based Stylized Radiance Fields for Dynamic Scenes
Xingyi Li, Zhiguo Cao, Yizheng Wu, Kewei Wang, Ke Xian, Zhe Wang, Guosheng Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20102-20112
Current 3D stylization methods often assume static scenes which violates the dynamic nature of our real world. To address this limitation we present S-DyRF a reference-based spatio-temporal stylization method for dynamic neural radiance fields. However stylizing dynamic 3D scenes is inherently challenging due to the limited availability of stylized reference images along the temporal axis. Our key insight lies in introducing additional temporal cues besides the provided reference. To this end we generate temporal pseudo-references from the given stylized reference. These pseudo-references facilitate the propagation of style information from the reference to the entire dynamic 3D scene. For coarse style transfer we enforce novel views and times to mimic the style details present in pseudo-references at the feature level. To preserve high-frequency details we create a collection of stylized temporal pseudo-rays from temporal pseudo-references. These pseudo-rays serve as detailed and explicit stylization guidance for achieving fine style transfer. Experiments on both synthetic and real-world datasets demonstrate that our method yields plausible stylized results of space-time view synthesis on dynamic 3D scenes.
****
MotionEditor: Editing Video Motion via Content-Aware Diffusion
Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7882-7891
Existing diffusion-based video editing models have made gorgeous advances for editing attributes of a source video over time but struggle to manipulate the motion information while preserving the original protagonist's appearance and background. To address this we propose MotionEditor the first diffusion model for video motion editing. MotionEditor incorporates a novel content-aware motion adapter into ControlNet to capture temporal motion correspondence. While ControlNet enables direct generation based on skeleton poses it encounters challenges when modifying the source motion in the inverted noise due to contradictory signals between the noise (source) and the condition (reference). Our adapter complements ControlNet by involving source content to transfer adapted control signals seamlessly. Further we build up a two-branch architecture (a reconstruction branch and an editing branch) with a high-fidelity attention injection mechanism facilitating branch interaction. This mechanism enables the editing branch to query the key and value from the reconstruction branch in a decoupled manner making the editing branch retain the original background and protagonist appearance. We also propose a skeleton alignment algorithm to address the discrepancies in pose size and position. Experiments demonstrate the promising motion editing ability of MotionEditor both qualitatively and quantitatively. To the best of our knowledge MotionEditor is the first to use diffusion models specifically for video motion editing considering the origin dynamic background and camera movement.

****
What How and When Should Object Detectors Update in Continually Changing Test Do
mains?
Jayeon Yoo, Dongkwan Lee, Inseop Chung, Donghyun Kim, Nojun Kwak; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 23354-23363
It is a well-known fact that the performance of deep learning models deteriorate
s when they encounter a distribution shift at test time. Test-time adaptation (T
TA) algorithms have been proposed to adapt the model online while inferring test
 data. However existing research predominantly focuses on classification tasks t
hrough the optimization of batch normalization layers or classification heads bu
t this approach limits its applicability to various model architectures like Tra
nsformers and makes it challenging to apply to other tasks such as object detect
ion. In this paper we propose a novel online adaption approach for object detect
ion in continually changing test domains considering which part of the model to
update how to update it and when to perform the update. By introducing architect
ure-agnostic and lightweight adaptor modules and only updating these while leavi
ng the pre-trained backbone unchanged we can rapidly adapt to new test domains i
n an efficient way and prevent catastrophic forgetting. Furthermore we present a
 practical and straightforward class-wise feature aligning method for object det
ection to resolve domain shifts. Additionally we enhance efficiency by determini
ng when the model is sufficiently adapted or when additional adaptation is neede
d due to changes in the test distribution. Our approach surpasses baselines on w
idely used benchmarks achieving improvements of up to 4.9%p and 7.9%p in mAP for
 COCO ? COCO-corrupted and SHIFT respectively while maintaining about 20 FPS or
higher. The implementation code is available at https://github.com/natureyoo/Con
tinualTTA_ObjectDetection.
****
One-Prompt to Segment All Medical Images
Junde Wu, Min Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2024, pp. 11302-11312
Large foundation models known for their strong zero-shot generalization have exc
elled in visual and language applications. However applying them to medical imag
e segmentation a domain with diverse imaging types and target labels remains an
open challenge. Current approaches such as adapting interactive segmentation mod
els like Segment Anything Model (SAM) require user prompts for each sample durin
g inference. Alternatively transfer learning methods like few/one-shot models de
mand labeled samples leading to high costs. This paper introduces a new paradigm
 toward the universal medical image segmentation termed 'One-Prompt Segmentation
.' One-Prompt Segmentation combines the strengths of one-shot and interactive me
thods. In the inference stage with just one prompted sample it can adeptly handl
e the unseen task in a single forward pass. We train One-Prompt Model on 64 open
-source medical datasets accompanied by the collection of over 3000 clinician-la
beled prompts. Tested on 14 previously unseen datasets the One-Prompt Model show
cases superior zero-shot segmentation capabilities outperforming a wide range of
 related methods. The code and data is released as https://github.com/KidsWithTo
kens/one-prompt.
****
Bayesian Exploration of Pre-trained Models for Low-shot Image Classification
Yibo Miao, Yu Lei, Feng Zhou, Zhijie Deng; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23849-23859
Low-shot image classification is a fundamental task in computer vision and the e
mergence of large-scale vision-language models such as CLIP has greatly advanced
 the forefront of research in this field. However most existing CLIP-based metho
ds lack the flexibility to effectively incorporate other pre-trained models that
 encompass knowledge distinct from CLIP. To bridge the gap this work proposes a
simple and effective probabilistic model ensemble framework based on Gaussian pr
ocesses which have previously demonstrated remarkable efficacy in processing sma
ll data. We achieve the integration of prior knowledge by specifying the mean fu
nction with CLIP and the kernel function with an ensemble of deep kernels built

upon various pre-trained models. By regressing the classification label directly our framework enables analytical inference straightforward uncertainty quantification and principled hyper-parameter tuning. Through extensive experiments on standard benchmarks we demonstrate that our method consistently outperforms competitive ensemble baselines regarding predictive performance. Additionally we assess the robustness of our method and the quality of the yielded uncertainty estimates on out-of-distribution datasets. We also illustrate that our method despite relying on label regression still enjoys superior model calibration compared to most deterministic baselines.
****
GROUNDHOG: Grounding Large Language Models to Holistic Segmentation
Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, Joyce Chai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14227-14238
Most multimodal large language models (MLLMs) learn language-to-object grounding through causal language modeling where grounded objects are captured by bounding boxes as sequences of location tokens. This paradigm lacks pixel-level representations that are important for fine-grained visual understanding and diagnosis. In this work we introduce GROUNDHOG an MLLM developed by grounding Large Language Models to holistic segmentation. GROUNDHOG incorporates a masked feature extractor and converts extracted features into visual entity tokens for the MLLM backbone which then connects groundable phrases to unified grounding masks by retrieving and merging the entity masks. To train GROUNDHOG we carefully curated M3G2 a grounded visual instruction tuning dataset with Multi-Modal Multi-Grained Grounding by harvesting a collection of segmentation-grounded datasets with rich annotations. Our experimental results show that GROUNDHOG achieves superior performance on various language grounding tasks without task-specific fine-tuning and significantly reduces object hallucination. GROUNDHOG also demonstrates better grounding towards complex forms of visual input and provides easy-to-understand diagnosis in failure cases.
****
Doubly Abductive Counterfactual Inference for Text-based Image Editing
Xue Song, Jiequan Cui, Hanwang Zhang, Jingjing Chen, Richang Hong, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9162-9171
We study text-based image editing (TBIE) of a single image by counterfactual inference because it is an elegant formulation to precisely address the requirement: the edited image should retain the fidelity of the original one. Through the lens of the formulation we find that the crux of TBIE is that existing techniques hardly achieve a good trade-off between editability and fidelity mainly due to the overfitting of the single-image fine-tuning. To this end we propose a Doubly Abductive Counterfactual inference framework (DAC). We first parameterize an exogenous variable as a UNet LoRA whose abduction can encode all the image details. Second we abduct another exogenous variable parameterized by a text encoder LoRA which recovers the lost editability caused by the overfitted first abduction. Thanks to the second abduction which exclusively encodes the visual transition from post-edit to pre-edit its inversion---subtracting the LoRA---effectively reverts pre-edit back to post-edit thereby accomplishing the edit. Through extensive experiments our DAC achieves a good trade-off between editability and fidelity. Thus we can support a wide spectrum of user editing intents including addition removal manipulation replacement style transfer and facial change which are extensively validated in both qualitative and quantitative evaluations. Codes are in https://github.com/xuesong39/DAC.
****
RoMa: Robust Dense Feature Matching
Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, Michael Felsberg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19790-19800
Feature matching is an important computer vision task that involves estimating correspondences between two images of a 3D scene and dense methods estimate all s

uch correspondences. The aim is to learn a robust model i.e. a model able to match under challenging real-world changes. In this work we propose such a model leveraging frozen pretrained features from the foundation model DINOv2. Although these features are significantly more robust than local features trained from scratch they are inherently coarse. We therefore combine them with specialized ConvNet fine features creating a precisely localizable feature pyramid. To further improve robustness we propose a tailored transformer match decoder that predicts anchor probabilities which enables it to express multimodality. Finally we propose an improved loss formulation through regression-by-classification with subsequent robust regression. We conduct a comprehensive set of experiments that show that our method RoMa achieves significant gains setting a new state-of-the-art. In particular we achieve a 36% improvement on the extremely challenging WxBS benchmark. Code is provided at github.com/Parskatt/RoMa.
****
Omni-SMoLA: Boosting Generalist Multimodal Models with Soft Mixture of Low-rank Experts
Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, Radu Soricut; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14205-14215
In this work we present Omni-SMoLA a multimodal architecture that mixes many multi-modal experts efficiently and achieves both high specialist and generalist performance. In contrast to previous models for which we see performance degradation on average when training the models on a wide range of tasks we show that the SMoLA low-rank experts are able to model different skills and task and overall improve the performance of a generalist model. This finding indicates that simple LMM fine-tuning is suboptimal for handling a wide range of tasks and that pairing the act of fine-tuning with specifically-designed architecture changes leads to better performing models.
****
SeMoLi: What Moves Together Belongs Together
Jenny Seidenschwarz, Aljosa Osep, Francesco Ferroni, Simon Lucey, Laura Leal-Taixe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14685-14694
We tackle semi-supervised object detection based on motion cues. Recent results suggest that heuristic-based clustering methods in conjunction with object trackers can be used to pseudo-label instances of moving objects and use these as supervisory signals to train 3D object detectors in Lidar data without manual supervision. We re-think this approach and suggest that both object detection as well as motion-inspired pseudo-labeling can be tackled in a data-driven manner. We leverage recent advances in scene flow estimation to obtain point trajectories from which we extract long-term class-agnostic motion patterns. Revisiting correlation clustering in the context of message passing networks we learn to group those motion patterns to cluster points to object instances. By estimating the full extent of the objects we obtain per-scan 3D bounding boxes that we use to supervise a Lidar object detection network. Our method not only outperforms prior heuristic-based approaches (57.5 AP +14 improvement over prior work) more importantly we show we can pseudo-label and train object detectors across datasets.
****
Insights from the Use of Previously Unseen Neural Architecture Search Datasets
Rob Geada, David Towers, Matthew Forshaw, Amir Atapour-Abarghouei, A. Stephen McGough; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22541-22550
The boundless possibility of neural networks which can be used to solve a problem - each with different performance - leads to a situation where a Deep Learning expert is required to identify the best neural network. This goes against the hope of removing the need for experts. Neural Architecture Search (NAS) offers a solution to this by automatically identifying the best architecture. However to date NAS work has focused on a small set of datasets which we argue are not representative of real-world problems. We introduce eight new datasets created for a series of NAS Challenges: AddNIST Language MultNIST CIFARTile Gutenberg Isabell

a GeoClassing and Chesseract. These datasets and challenges are developed to direct attention to issues in NAS development and to encourage authors to consider how their models will perform on datasets unknown to them at development time. We present experimentation using standard Deep Learning methods as well as the best results from challenge participants
****
Adversarially Robust Few-shot Learning via Parameter Co-distillation of Similarity and Class Concept Learners
Junhao Dong, Piotr Koniusz, Junxi Chen, Xiaohua Xie, Yew-Soon Ong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28535-28544
Few-shot learning (FSL) facilitates a variety of computer vision tasks yet remains vulnerable to adversarial attacks. Existing adversarially robust FSL methods rely on either visual similarity learning or class concept learning. Our analysis reveals that these two learning paradigms are complementary exhibiting distinct robustness due to their unique decision boundary types (concepts clustering by the visual similarity label vs. classification by the class labels). To bridge this gap we propose a novel framework unifying adversarially robust similarity learning and class concept learning. Specifically we distill parameters from both network branches into a "unified embedding model" during robust optimization and redistribute them to individual network branches periodically. To capture generalizable robustness across diverse branches we initialize adversaries in each episode with cross-branch class-wise "global adversarial perturbations" instead of less informative random initialization. We also propose a branch robustness harmonization to modulate the optimization of similarity and class concept learners via their relative adversarial robustness. Extensive experiments demonstrate the state-of-the-art performance of our method in diverse few-shot scenarios.
****
Context-Guided Spatio-Temporal Video Grounding
Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, Libo Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18330-18339
Spatio-temporal video grounding (or STVG) task aims at locating a spatio-temporal tube for a specific instance given a text query. Despite advancements current methods easily suffer the distractors or heavy object appearance variations in videos due to insufficient object information from the text leading to degradation. Addressing this we propose a novel framework context-guided STVG (CG-STVG) which mines discriminative instance context for object in videos and applies it as a supplementary guidance for target localization. The key of CG-STVG lies in two specially designed modules including instance context generation (ICG) which focuses on discovering visual context information (in both appearance and motion) of the instance and instance context refinement (ICR) which aims to improve the instance context from ICG by eliminating irrelevant or even harmful information from the context. During grounding ICG together with ICR are deployed at each decoding stage of a Transformer architecture for instance context learning. Particularly instance context learned from one decoding stage is fed to the next stage and leveraged as a guidance containing rich and discriminative object feature to enhance the target-awareness in decoding feature which conversely benefits generating better new instance context for improving localization finally. Compared to existing methods CG-STVG enjoys object information in text query and guidance from mined instance visual context for more accurate target localization. In our experiments on three benchmarks including HCSTVG-v1/-v2 and VidSTG CG-STVG sets new state-of-the-arts in m_tIoU and m_vIoU on all of them showing efficacy. Code is released at https://github.com/HengLan/CGSTVG.
****
Explaining the Implicit Neural Canvas: Connecting Pixels to Neurons by Tracing their Contributions
Namitha Padmanabhan, Matthew Gwilliam, Pulkit Kumar, Shishira R Maiya, Max Ehrlich, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10957-10967

The many variations of Implicit Neural Representations (INRs) where a neural net
work is trained as a continuous representation of a signal have tremendous pract
ical utility for downstream tasks including novel view synthesis video compressi
on and image super-resolution. Unfortunately the inner workings of these network
s are seriously understudied. Our work eXplaining the Implicit Neural Canvas (XI
NC) is a unified framework for explaining properties of INRs by examining the st
rength of each neuron's contribution to each output pixel. We call the aggregate
 of these contribution maps the Implicit Neural Canvas and we use this concept t
o demonstrate that the INRs we study learn to "see" the frames they represent in
 surprising ways. For example INRs tend to have highly distributed representatio
ns. While lacking high-level object semantics they have a significant bias for c
olor and edges and are almost entirely space-agnostic. We arrive at our conclusi
ons by examining how objects are represented across time in video INRs using clu
stering to visualize similar neurons across layers and architectures and show th
at this is dominated by motion. These insights demonstrate the general usefulnes
s of our analysis framework.
****
APISR: Anime Production Inspired Real-World Anime Super-Resolution
Boyang Wang, Fengyu Yang, Xihang Yu, Chao Zhang, Hanbin Zhao; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 25574-25584
While real-world anime super-resolution (SR) has gained increasing attention in
the SR community existing methods still adopt techniques from the photorealistic
 domain. In this paper we analyze the anime production workflow and rethink how
to use characteristics of it for the sake of the real-world anime SR. First we a
rgue that video networks and datasets are not necessary for anime SR due to the
repetition use of hand-drawing frames. Instead we propose an anime image collect
ion pipeline by choosing the least compressed and the most informative frames fr
om the video sources. Based on this pipeline we introduce the Anime Production-o
riented Image (API) dataset. In addition we identify two anime-specific challeng
es of distorted and faint hand-drawn lines and unwanted color artifacts. We addr
ess the first issue by introducing a prediction-oriented compression module in t
he image degradation model and a pseudo-ground truth preparation with enhanced h
and-drawn lines. In addition we introduce the balanced twin perceptual loss comb
ining both anime and photorealistic high-level features to mitigate unwanted col
or artifacts and increase visual clarity. We evaluate our method through extensi
ve experiments on the public benchmark showing our method outperforms state-of-t
he-art anime dataset-trained approaches.
****
MVCPS-NeuS: Multi-view Constrained Photometric Stereo for Neural Surface Reconst
ruction
Hiroaki Santo, Fumio Okura, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20475-20484
Multi-view photometric stereo (MVPS) recovers a high-fidelity 3D shape of a scen
e by benefiting from both multi-view stereo and photometric stereo. While photom
etric stereo boosts detailed shape reconstruction it necessitates recording imag
es under various light conditions for each viewpoint. In particular calibrating
the light directions for each view significantly increases the cost of acquiring
 images. To make MVPS more accessible we introduce a practical and easy-to-imple
ment setup multi-view constrained photometric stereo (MVCPS) where the light dir
ections are unknown but constrained to move together with the camera. Unlike con
ventional multi-view uncalibrated photometric stereo our constrained setting red
uces the ambiguities of surface normal estimates from per-view linear ambiguitie
s to a single and global linear one thereby simplifying the disambiguation proce
ss. The proposed method integrates the ambiguous surface normal into neural surf
ace reconstruction (NeuS) to simultaneously resolve the global ambiguity and est
imate the detailed 3D shape. Experiments demonstrate that our method estimates a
ccurate shapes under sparse viewpoints using only a few multi-view constrained l
ight sources.
****

ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding
Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, Silvio Savarese; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27091-27101
Recent advancements in multimodal pre-training have shown promising efficacy in 3D representation learning by aligning multimodal features across 3D shapes their 2D counterparts and language descriptions. However the methods used by existing frameworks to curate such multimodal data in particular language descriptions for 3D shapes are not scalable and the collected language descriptions are not diverse. To address this we introduce ULIP-2 a simple yet effective tri-modal pre training framework that leverages large multimodal models to automatically generate holistic language descriptions for 3D shapes. It only needs 3D data as input eliminating the need for any manual 3D annotations and is therefore scalable to large datasets. ULIP-2 is also equipped with scaled-up backbones for better multi-modal representation learning. We conduct experiments on two large-scale 3D datasets Objaverse and ShapeNet and augment them with tri-modal datasets of 3D point clouds images and language for training ULIP-2. Experiments show that ULIP-2 demonstrates substantial benefits in three downstream tasks: zero-shot 3D classification standard 3D classification with fine-tuning and 3D captioning (3D-to-language generation). It achieves a new SOTA of 50.6% (top- 1) on Objaverse-LVIS and 84.7% (top-1) on ModelNet40 in zero-shot classification. In the ScanObjectNN benchmark for standard fine-tuning ULIP-2 reaches an overall accuracy of 91.5% with a compact model of only 1.4 million parameters. ULIP-2 sheds light on a new paradigm for scalable multimodal 3D representation learning without human annotations and shows significant improvements over existing baselines. The code and datasets are released at https://github.com/salesforce/ULIP.
****
Normalizing Flows on the Product Space of SO(3) Manifolds for Probabilistic Human Pose Modeling
Olaf Dünkel, Tim Salzmann, Florian Pfaff; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2285-2294
Normalizing flows have proven their efficacy for density estimation in Euclidean space but their application to rotational representations crucial in various domains such as robotics or human pose modeling remains underexplored. Probabilistic models of the human pose can benefit from approaches that rigorously consider the rotational nature of human joints. For this purpose we introduce HuProSO3 a normalizing flow model that operates on a high-dimensional product space of SO(3) manifolds modeling the joint distribution for human joints with three degrees of freedom. HuProSO3's advantage over state-of-the-art approaches is demonstrated through its superior modeling accuracy in three different applications and its capability to evaluate the exact likelihood. This work not only addresses the technical challenge of learning densities on SO(3) manifolds but it also has broader implications for domains where the probabilistic regression of correlated 3D rotations is of importance. Code will be available at https://github.com/odunkel/HuProSO.
****
Adapting to Length Shift: FlexiLength Network for Trajectory Prediction
Yi Xu, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15226-15237
Trajectory prediction plays an important role in various applications including autonomous driving robotics and scene understanding. Existing approaches mainly focus on developing compact neural networks to increase prediction precision on public datasets typically employing a standardized input duration. However a notable issue arises when these models are evaluated with varying observation lengths leading to a significant performance drop a phenomenon we term the Observation Length Shift. To address this issue we introduce a general and effective framework the FlexiLength Network (FLN) to enhance the robustness of existing trajectory prediction techniques against varying observation periods. Specifically FLN integrates trajectory data with diverse observation lengths incorporates FlexiLe

ngth Calibration (FLC) to acquire temporal invariant representations and employs FlexiLength Adaptation (FLA) to further refine these representations for more accurate future trajectory predictions. Comprehensive experiments on multiple datasets i.e. ETH/UCY nuScenes and Argoverse 1 demonstrate the effectiveness and flexibility of our proposed FLN framework.
****
WorDepth: Variational Language Prior for Monocular Depth Estimation
Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Stefano Soatto, Dong Lao, Alex Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9708-9719
Three-dimensional (3D) reconstruction from a single image is an ill-posed problem with inherent ambiguities i.e. scale. Predicting a 3D scene from text description(s) is similarly ill-posed i.e. spatial arrangements of objects described. We investigate the question of whether two inherently ambiguous modalities can be used in conjunction to produce metric-scaled reconstructions. To test this we focus on monocular depth estimation the problem of predicting a dense depth map from a single image but with an additional text caption describing the scene. To this end we begin by encoding the text caption as a mean and standard deviation; using a variational framework we learn the distribution of the plausible metric reconstructions of 3D scenes corresponding to the text captions as a prior. To "select" a specific reconstruction or depth map we encode the given image through a conditional sampler that samples from the latent space of the variational text encoder which is then decoded to the output depth map. Our approach is trained alternatingly between the text and image branches: in one optimization step we predict the mean and standard deviation from the text description and sample from a standard Gaussian and in the other we sample using a (image) conditional sampler. Once trained we directly predict depth from the encoded text using the conditional sampler. We demonstrate our approach on indoor (NYUv2) and outdoor (KITTI) scenarios where we show that language can consistently improve performance in both. Code: https://github.com/Adonis-galaxy/WorDepth.
****
WaveMo: Learning Wavefront Modulations to See Through Scattering
Mingyang Xie, Haiyun Guo, Brandon Y. Feng, Lingbo Jin, Ashok Veeraraghavan, Christopher A. Metzler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25276-25285
Imaging through scattering media is a fundamental and pervasive challenge in fields ranging from medical diagnostics to astronomy. A promising strategy to overcome this challenge is wavefront modulation which induces measurement diversity during image acquisition. Despite its importance designing optimal wavefront modulations to image through scattering remains under-explored. This paper introduces a novel learning-based framework to address the gap. Our approach jointly optimizes wavefront modulations and a computationally lightweight feedforward "proxy" reconstruction network. This network is trained to recover scenes obscured by scattering using measurements that are modified by these modulations. The learned modulations produced by our framework generalize effectively to unseen scattering scenarios and exhibit remarkable versatility. During deployment the learned modulations can be decoupled from the proxy network to augment other more computationally expensive restoration algorithms. Through extensive experiments we demonstrate our approach significantly advances the state of the art in imaging through scattering media. Our project webpage is at https://wavemo-2024.github.io/.
****
ReGenNet: Towards Human Action-Reaction Synthesis
Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, Wenjun Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1759-1769
Humans constantly interact with their surrounding environments. Current human-centric generative models mainly focus on synthesizing humans plausibly interacting with static scenes and objects while the dynamic human action-reaction synthesis for ubiquitous causal human-human interactions is less explored. Human-human interactions can be regarded as asymmetric with actors and reactors in atomic in

teraction periods. In this paper we comprehensively analyze the asymmetric dynamic synchronous and detailed nature of human-human interactions and propose the first multi-setting human action-reaction synthesis benchmark to generate human reactions conditioned on given human actions. To begin with we propose to annotate the actor-reactor order of the interaction sequences for the NTU120 InterHuman and Chi3D datasets. Based on them a diffusion-based generative model with a Transformer decoder architecture called ReGenNet together with an explicit distance-based interaction loss is proposed to predict human reactions in an online manner where the future states of actors are unavailable to reactors. Quantitative and qualitative results show that our method can generate instant and plausible human reactions compared to the baselines and can generalize to unseen actor motions and viewpoint changes.
****
A Simple Baseline for Efficient Hand Mesh Reconstruction
Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, Jiajun Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1367-1376
Hand mesh reconstruction has attracted considerable attention in recent years with various approaches and techniques being proposed. Some of these methods incorporate complex components and designs which while effective may complicate the model and hinder efficiency. In this paper we decompose the mesh decoder into token generator and mesh regressor. Through extensive ablation experiments we found that the token generator should select discriminating and representative points while the mesh regressor needs to upsample sparse keypoints into dense meshes in multiple stages. Given these functionalities we can achieve high performance with minimal computational resources. Based on this observation we propose a simple yet effective baseline that outperforms state-of-the-art methods by a large margin while maintaining real-time efficiency. Our method outperforms existing solutions achieving state-of-the-art (SOTA) results across multiple datasets. On the FreiHAND dataset our approach produced a PA-MPJPE of 5.8mm and a PA-MPVPE of 6.1mm. Similarly on the DexYCB dataset we observed a PA-MPJPE of 5.5mm and a PA-MPVPE of 5.5mm. As for performance speed our method reached up to 33 frames per second (fps) when using HRNet and up to 70 fps when employing FastViT-MA36. Code will be made available.
****
Integrating Efficient Optimal Transport and Functional Maps For Unsupervised Shape Correspondence Learning
Tung Le, Khai Nguyen, Shanlin Sun, Nhat Ho, Xiaohui Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23188-23198
In the realm of computer vision and graphics accurately establishing correspondences between geometric 3D shapes is pivotal for applications like object tracking registration texture transfer and statistical shape analysis. Moving beyond traditional hand-crafted and data-driven feature learning methods we incorporate spectral methods with deep learning focusing on functional maps (FMs) and optimal transport (OT). Traditional OT-based approaches often reliant on entropy regularization OT in learning-based framework face computational challenges due to their quadratic cost. Our key contribution is to employ the sliced Wasserstein distance (SWD) for OT which is a valid fast optimal transport metric in an unsupervised shape matching framework. This unsupervised framework integrates functional map regularizers with a novel OT-based loss derived from SWD enhancing feature alignment between shapes treated as discrete probability measures. We also introduce an adaptive refinement process utilizing entropy regularized OT further refining feature alignments for accurate point-to-point correspondences. Our method demonstrates superior performance in non-rigid shape matching including near-isometric and non-isometric scenarios and excels in downstream tasks like segmentation transfer. The empirical results on diverse datasets highlight our framework's effectiveness and generalization capabilities setting new standards in non-rigid shape matching with efficient OT metrics and an adaptive refinement module.
****

PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding
Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8640-8650
Recent advances in text-to-image generation have made remarkable progress in synthesizing realistic human photos conditioned on given text prompts. However existing personalized generation methods cannot simultaneously satisfy the requirements of high efficiency promising identity (ID) fidelity and flexible text controllability. In this work we introduce PhotoMaker an efficient personalized text-to-image generation method which mainly encodes an arbitrary number of input ID images into a stack ID embedding for preserving ID information. Such an embedding also empowers our method to be applied in many interesting scenarios such as when replacing the corresponding class word and when combining the characteristics of different identities. Besides to better drive the training of our PhotoMaker we propose an ID-oriented data creation pipeline to assemble the training data. Under the nourishment of the dataset constructed through the proposed pipeline our PhotoMaker demonstrates comparable performance to test-time fine-tuning-based methods yet provides significant speed improvements strong generalization capabilities and a wide range of applications.
****
Score-Guided Diffusion for 3D Human Recovery
Anastasis Stathopoulos, Ligong Han, Dimitris Metaxas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 906-915
We present Score-Guided Human Mesh Recovery (ScoreHMR) an approach for solving inverse problems for 3D human pose and shape reconstruction. These inverse problems involve fitting a human body model to image observations traditionally solved through optimization techniques. ScoreHMR mimics model fitting approaches but alignment with the image observation is achieved through score guidance in the latent space of a diffusion model. The diffusion model is trained to capture the conditional distribution of the human model parameters given an input image. By guiding its denoising process with a task-specific score ScoreHMR effectively solves inverse problems for various applications without the need for retraining the task-agnostic diffusion model. We evaluate our approach on three settings/applications. These are: (i) single-frame model fitting; (ii) reconstruction from multiple uncalibrated views; (iii) reconstructing humans in video sequences. ScoreHMR consistently outperforms all optimization baselines on popular benchmarks across all settings. We make our code and models available on the project website: https://statho.github.io/ScoreHMR.
****
Check Locate Rectify: A Training-Free Layout Calibration System for Text-to-Image Generation
Biao Gong, Siteng Huang, Yutong Feng, Shiwei Zhang, Yuyuan Li, Yu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6624-6634
Diffusion models have recently achieved remarkable progress in generating realistic images. However challenges remain in accurately understanding and synthesizing the layout requirements in the textual prompts. To align the generated image with layout instructions we present a training-free layout calibration system SimM that intervenes in the generative process on the fly during inference time. Specifically following a "check-locate-rectify" pipeline the system first analyses the prompt to generate the target layout and compares it with the intermediate outputs to automatically detect errors. Then by moving the located activations and making intra- and inter-map adjustments the rectification process can be performed with negligible computational overhead. To evaluate SimM over a range of layout requirements we present a benchmark SimMBench that compensates for the lack of superlative spatial relations in existing datasets. And both quantitative and qualitative results demonstrate the effectiveness of the proposed SimM in calibrating the layout inconsistencies. Our project page is at https://simm-t2i.github.io/SimM.

****
ODCR: Orthogonal Decoupling Contrastive Regularization for Unpaired Image Dehazing
Zhongze Wang, Haitao Zhao, Jingchao Peng, Lujian Yao, Kaijie Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25479-25489
Unpaired image dehazing (UID) holds significant research importance due to the challenges in acquiring haze/clear image pairs with identical backgrounds. This paper proposes a novel method for UID named Orthogonal Decoupling Contrastive Regularization (ODCR). Our method is grounded in the assumption that an image consists of both haze-related features which influence the degree of haze and haze-unrelated features such as texture and semantic information. ODCR aims to ensure that the haze-related features of the dehazing result closely resemble those of the clear image while the haze-unrelated features align with the input hazy image. To accomplish the motivation Orthogonal MLPs optimized geometrically on the Stiefel manifold are proposed which can project image features into an orthogonal space thereby reducing the relevance between different features. Furthermore a task-driven Depth-wise Feature Classifier (DWFC) is proposed which assigns weights to the orthogonal features based on the contribution of each channel's feature in predicting whether the feature source is hazy or clear in a self-supervised fashion. Finally a Weighted PatchNCE (WPNCE) loss is introduced to achieve the pulling of haze-related features in the output image toward those of clear images while bringing haze-unrelated features close to those of the hazy input. Extensive experiments demonstrate the superior performance of our ODCR method on UID.
****
Pose-Transformed Equivariant Network for 3D Point Trajectory Prediction
Ruixuan Yu, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5503-5512
Predicting 3D point trajectory is a fundamental learning task which commonly should be equivariant under Euclidean transformation e.g. SE(3). The existing equivariant models are commonly based on the group equivariant convolution equivariant message passing vector neuron frame averaging etc. In this paper we propose a novel pose-transformed equivariant network in which the points are firstly uniquely normalized and then transformed by the learned pose transformations upon which the points after motion are predicted and aggregated. Under each transformed pose we design the point position predictor consisting of multiple Pose-Transformed Points Prediction blocks in which the global and local motions are estimated and aggregated. This framework can be proven to be equivariant to SE(3) transformation over 3D points. We evaluate the pose-transformed equivariant network on extensive datasets including human motion capture molecular dynamics modeling and dynamics simulation. Extensive experimental comparisons demonstrated our SOTA performance compared with the existing equivariant networks for 3D point trajectory prediction.
****
OmniSeg3D: Omniversal 3D Segmentation via Hierarchical Contrastive Learning
Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, Lu Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20612-20622
Towards holistic understanding of 3D scenes a general 3D segmentation method is needed that can segment diverse objects without restrictions on object quantity or categories while also reflecting the inherent hierarchical structure. To achieve this we propose OmniSeg3D an omniversal segmentation method aims for segmenting anything in 3D all at once. The key insight is to lift multi-view inconsistent 2D segmentations into a consistent 3D feature field through a hierarchical contrastive learning framework which is accomplished by two steps. Firstly we design a novel hierarchical representation based on category-agnostic 2D segmentations to model the multi-level relationship among pixels. Secondly image features rendered from the 3D feature field are clustered at different levels which can be further drawn closer or pushed apart according to the hierarchical relationship between different levels. In tackling the challenges posed by inconsistent 2D s

egmentations this framework yields a global consistent 3D feature field which further enables hierarchical segmentation multi-object selection and global discretization. Extensive experiments demonstrate the effectiveness of our method on high-quality 3D segmentation and accurate hierarchical structure understanding. A graphical user interface further facilitates flexible interaction for omniversal 3D segmentation.
****
Revisiting Sampson Approximations for Geometric Estimation Problems
Felix Rydell, Angélica Torres, Viktor Larsson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4990-4998
Many problems in computer vision can be formulated as geometric estimation problems i.e. given a collection of measurements (e.g. point correspondences) we wish to fit a model (e.g. an essential matrix) that agrees with our observations. This necessitates some measure of how much an observation "agrees" with a given model. A natural choice is to consider the smallest perturbation that makes the observation exactly satisfy the constraints. However for many problems this metric is expensive or otherwise intractable to compute. The so-called Sampson error approximates this geometric error through a linearization scheme. For epipolar geometry the Sampson error is a popular choice and in practice known to yield very tight approximations of the corresponding geometric residual (the reprojection error). In this paper we revisit the Sampson approximation and provide new theoretical insights as to why and when this approximation works as well as provide explicit bounds on the tightness under some mild assumptions. Our theoretical results are validated in several experiments on real data and in the context of different geometric estimation tasks.
****
Fixed Point Diffusion Models
Xingjian Bai, Luke Melas-Kyriazi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9430-9440
We introduce the Fixed Point Diffusion Model (FPDM) a novel approach to image generation that integrates the concept of fixed point solving into the framework of diffusion-based generative modeling. Our approach embeds an implicit fixed point solving layer into the denoising network of a diffusion model transforming the diffusion process into a sequence of closely-related fixed point problems. Combined with a new stochastic training method this approach significantly reduces model size reduces memory usage and accelerates training. Moreover it enables the development of two new techniques to improve sampling efficiency: reallocating computation across timesteps and reusing fixed point solutions between timesteps. We conduct extensive experiments with state-of-the-art models on ImageNet FFHQ CelebA-HQ and LSUN-Church demonstrating substantial improvements in performance and efficiency. Compared to the state-of-the-art DiT model FPDM contains 87% fewer parameters consumes 60% less memory during training and improves image generation quality in situations where sampling computation or time is limited.
****
Simple Semantic-Aided Few-Shot Learning
Hai Zhang, Junzhe Xu, Shanlin Jiang, Zhenan He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28588-28597
Learning from a limited amount of data namely Few-Shot Learning stands out as a challenging computer vision task. Several works exploit semantics and design complicated semantic fusion mechanisms to compensate for rare representative features within restricted data. However relying on naive semantics such as class names introduces biases due to their brevity while acquiring extensive semantics from external knowledge takes a huge time and effort. This limitation severely constrains the potential of semantics in Few-Shot Learning. In this paper we design an automatic way called Semantic Evolution to generate high-quality semantics. The incorporation of high-quality semantics alleviates the need for complex network structures and learning algorithms used in previous works. Hence we employ a simple two-layer network termed Semantic Alignment Network to transform semantics and visual features into robust class prototypes with rich discriminative features for few-shot classification. The experimental results show our framework ou

tperforms all previous methods on six benchmarks demonstrating a simple network with high-quality semantics can beat intricate multi-modal modules on few-shot classification tasks. Code is available at https://github.com/zhangdoudou123/SemFew.
****
A Unified Framework for Microscopy Defocus Deblur with Multi-Pyramid Transformer and Contrastive Learning
Yuelin Zhang, Pengyu Zheng, Wanquan Yan, Chengyu Fang, Shing Shin Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11125-11136
Defocus blur is a persistent problem in microscope imaging that poses harm to pathology interpretation and medical intervention in cell microscopy and microscope surgery. To address this problem a unified framework including the multi-pyramid transformer (MPT) and extended frequency contrastive regularization (EFCR) is proposed to tackle two outstanding challenges in microscopy deblur: longer attention span and data deficiency. The MPT employs an explicit pyramid structure at each network stage that integrates the cross-scale window attention (CSWA) the intra-scale channel attention (ISCA) and the feature-enhancing feed-forward network (FEFN) to capture long-range cross-scale spatial interaction and global channel context. The EFCR addresses the data deficiency problem by exploring latent deblur signals from different frequency bands. It also enables deblur knowledge transfer to learn cross-domain information from extra data improving deblur performance for labeled and unlabeled data. Extensive experiments and downstream task validation show the framework achieves state-of-the-art performance across multiple datasets. Project page: https://github.com/PieceZhang/MPT-CataBlur.
****
Frozen Feature Augmentation for Few-Shot Image Classification
Andreas Bär, Neil Houlsby, Mostafa Dehghani, Manoj Kumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16046-16057
Training a linear classifier or lightweight model on top of pretrained vision model outputs so-called 'frozen features' leads to impressive performance on a number of downstream few-shot tasks. Currently frozen features are not modified during training. On the other hand when networks are trained directly on images data augmentation is a standard recipe that improves performance with no substantial overhead. In this paper we conduct an extensive pilot study on few-shot image classification that explores applying data augmentations in the frozen feature space dubbed 'frozen feature augmentation (FroFA)' covering twenty augmentations in total. Our study demonstrates that adopting a deceptively simple pointwise FroFA such as brightness can improve few-shot performance consistently across three network architectures three large pretraining datasets and eight transfer datasets.
****
Residual Learning in Diffusion Models
Junyu Zhang, Daochang Liu, Eunbyung Park, Shichao Zhang, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7289-7299
Diffusion models (DMs) have achieved remarkable generative performance particularly with the introduction of stochastic differential equations (SDEs). Nevertheless a gap emerges in the model sampling trajectory constructed by reverse-SDE due to the accumulation of score estimation and discretization errors. This gap results in a residual in the generated images adversely impacting the image quality. To remedy this we propose a novel residual learning framework built upon a correction function. The optimized function enables to improve image quality via rectifying the sampling trajectory effectively. Importantly our framework exhibits transferable residual correction ability i.e. a correction function optimized for one pre-trained DM can also enhance the sampling trajectory constructed by other different DMs on the same dataset. Experimental results on four widely-used datasets demonstrate the effectiveness and transferable capability of our framework.

****
Leveraging Cross-Modal Neighbor Representation for Improved CLIP Classification
Chao Yi, Lu Ren, De-Chuan Zhan, Han-Jia Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27402-27411
CLIP showcases exceptional cross-modal matching capabilities due to its training on image-text contrastive learning tasks. However without specific optimization for unimodal scenarios its performance in single-modality feature extraction might be suboptimal. Despite this some studies have directly used CLIP's image encoder for tasks like few-shot classification introducing a misalignment between its pre-training objectives and feature extraction methods. This inconsistency can diminish the quality of the image's feature representation adversely affecting CLIP's effectiveness in target tasks. In this paper we view text features as precise neighbors of image features in CLIP's space and present a novel CrOss-moDal nEighbor Representation (CODER) based on the distance structure between images and their neighbor texts. This feature extraction method aligns better with CLIP's pre-training objectives thereby fully leveraging CLIP's robust cross-modal capabilities. The key to construct a high-quality CODER lies in how to create a vast amount of high-quality and diverse texts to match with images. We introduce the Auto Text Generator (ATG) to automatically produce the required text in a data-free and training-free manner. We apply CODER to CLIP's zero-shot and few-shot image classification tasks. Experiment results across various datasets and models confirm CODER's effectiveness. Code is available at: https://github.com/YCaigogogo/CVPR24-CODER.
****
Beyond Textual Constraints: Learning Novel Diffusion Conditions with Fewer Examples
Yuyang Yu, Bangzhen Liu, Chenxi Zheng, Xuemiao Xu, Huaidong Zhang, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7109-7118
In this paper we delve into a novel aspect of learning novel diffusion conditions with datasets an order of magnitude smaller. The rationale behind our approach is the elimination of textual constraints during the few-shot learning process. To that end we implement two optimization strategies. The first prompt-free conditional learning utilizes a prompt-free encoder derived from a pre-trained Stable Diffusion model. This strategy is designed to adapt new conditions to the diffusion process by minimizing the textual-visual correlation thereby ensuring a more precise alignment between the generated content and the specified conditions. The second strategy entails condition-specific negative rectification which addresses the inconsistencies typically brought about by Classifier-free guidance in few-shot training contexts. Our extensive experiments across a variety of condition modalities demonstrate the effectiveness and efficiency of our framework yielding results comparable to those obtained with datasets a thousand times larger. Our codes are available at https://github.com/Yuyan9Yu/BeyondTextConstraint.
****
Incorporating Geo-Diverse Knowledge into Prompting for Increased Geographical Robustness in Object Recognition
Kyle Buettner, Sina Malakouti, Xiang Lorraine Li, Adriana Kovashka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13515-13524
Existing object recognition models have been shown to lack robustness in diverse geographical scenarios due to domain shifts in design and context. Class representations need to be adapted to more accurately reflect an object concept under these shifts. In the absence of training data from target geographies we hypothesize that geographically diverse descriptive knowledge of categories can enhance robustness. For this purpose we explore the feasibility of probing a large language model for geography-based object knowledge and we examine the effects of integrating knowledge into zero-shot and learnable soft prompting with CLIP. Within this exploration we propose geography knowledge regularization to ensure that soft prompts trained on a source set of geographies generalize to an unseen targ

et set. Accuracy gains over prompting baselines on DollarStreet while training only on Europe data are up to +2.8/1.2/1.6 on target data from Africa/Asia/Americas and +4.6 overall on the hardest classes. Competitive performance is shown vs. few-shot target training and analysis is provided to direct future study of geographical robustness.
****
Revisiting Adversarial Training Under Long-Tailed Distributions
Xinli Yue, Ningping Mou, Qian Wang, Lingchen Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24492-24501
Deep neural networks are vulnerable to adversarial attacks leading to erroneous outputs. Adversarial training has been recognized as one of the most effective methods to counter such attacks. However existing adversarial training techniques have predominantly been evaluated on balanced datasets whereas real-world data often exhibit a long-tailed distribution casting doubt on the efficacy of these methods in practical scenarios. In this paper we delve into the performance of adversarial training under long-tailed distributions. Through an analysis of the prior method "RoBal" (Wu et al. CVPR'21) we discover that utilizing Balanced Softmax Loss (BSL) alone can obtain comparable performance to the complete RoBal approach while significantly reducing the training overhead. Then we reveal that adversarial training under long-tailed distributions also suffers from robust overfitting similar to uniform distributions. We explore utilizing data augmentation to mitigate this issue and unexpectedly discover that unlike results obtained with balanced data data augmentation not only effectively alleviates robust overfitting but also significantly improves robustness. We further identify that the improvement is attributed to the increased diversity of training data. Extensive experiments further corroborate that data augmentation alone can significantly improve robustness. Finally building on these findings we demonstrate that compared to RoBal the combination of BSL and data augmentation leads to a +6.66% improvement in model robustness under AutoAttack on CIFAR-10-LT. Our code is available at: https://github.com/NISPLab/AT-BSL.
****
Exploiting Style Latent Flows for Generalizing Deepfake Video Detection
Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, Jongwon Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1133-1143
This paper presents a new approach for the detection of fake videos based on the analysis of style latent vectors and their abnormal behavior in temporal changes in the generated videos. We discovered that the generated facial videos suffer from the temporal distinctiveness in the temporal changes of style latent vectors which are inevitable during the generation of temporally stable videos with various facial expressions and geometric transformations. Our framework utilizes the StyleGRU module trained by contrastive learning to represent the dynamic properties of style latent vectors. Additionally we introduce a style attention module that integrates StyleGRU-generated features with content-based features enabling the detection of visual and temporal artifacts. We demonstrate our approach across various benchmark scenarios in deepfake detection showing its superiority in cross-dataset and cross-manipulation scenarios. Through further analysis we also validate the importance of using temporal changes of style latent vectors to improve the generality of deepfake video detection.
****
PIN: Positional Insert Unlocks Object Localisation Abilities in VLMs
Michael Dorkenwald, Nimrod Barazani, Cees G. M. Snoek, Yuki M. Asano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13548-13558
Vision-Language Models (VLMs) such as Flamingo and GPT-4V have shown immense potential by integrating large language models with vision systems. Nevertheless these models face challenges in the fundamental computer vision task of object localisation due to their training on multimodal data containing mostly captions without explicit spatial grounding. While it is possible to construct custom super

vised training pipelines with bounding box annotations that integrate with VLMs these result in specialized and hard-to-scale models. In this paper we aim to explore the limits of caption-based VLMs and instead propose to tackle the challenge in a simpler manner by i) keeping the weights of a caption-based VLM frozen and ii) not using any supervised detection data. To this end we introduce an input-agnostic Positional Insert (PIN) a learnable spatial prompt containing a minimal set of parameters that are slid inside the frozen VLM unlocking object localisation capabilities. Our PIN module is trained with a simple next-token prediction task on synthetic data without requiring the introduction of new output heads. Our experiments demonstrate strong zero-shot localisation performances on a variety of images including Pascal VOC COCO LVIS and diverse images like paintings or cartoons.
****
UniGarmentManip: A Unified Framework for Category-Level Garment Manipulation via Dense Visual Correspondence
Ruihai Wu, Haoran Lu, Yiyan Wang, Yubo Wang, Hao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16340-16350
Garment manipulation (e.g. unfolding folding and hanging clothes) is essential for future robots to accomplish home-assistant tasks while highly challenging due to the diversity of garment configurations geometries and deformations. Although able to manipulate similar shaped garments in a certain task previous works mostly have to design different policies for different tasks could not generalize to garments with diverse geometries and often rely heavily on human-annotated data. In this paper we leverage the property that garments in a certain category have similar structures and then learn the topological dense (point-level) visual correspondence among garments in the category level with different deformations in the self-supervised manner. The topological correspondence can be easily adapted to the functional correspondence to guide the manipulation policies for various downstream tasks within only one or few-shot demonstrations. Experiments over garments in 3 different categories on 3 representative tasks in diverse scenarios using one or two arms taking one or more steps inputting flat or messy garments demonstrate the effectiveness of our proposed method. Project page: https://warshallrho.github.io/unigarmentmanip.
****
Multi-Attribute Interactions Matter for 3D Visual Grounding
Can Xu, Yuehui Han, Rui Xu, Le Hui, Jin Xie, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17253-17262
3D visual grounding aims to localize 3D objects described by free-form language sentences. Following the detection-then-matching paradigm existing methods mainly focus on embedding object attributes in unimodal feature extraction and multimodal feature fusion to enhance the discriminability of the proposal feature for accurate grounding. However most of them ignore the explicit interaction of multiple attributes causing a bias in unimodal representation and misalignment in multimodal fusion. In this paper we propose a multi-attribute aware Transformer for 3D visual grounding learning the multi-attribute interactions to refine the intra-modal and inter-modal grounding cues. Specifically we first develop an attribute causal analysis module to quantify the causal effect of different attributes for the final prediction which provides powerful supervision to correct the misleading attributes and adaptively capture other discriminative features. Then we design an exchanging-based multimodal fusion module which dynamically replaces tokens with low attribute attention between modalities before directly integrating low-dimensional global features. This ensures an attribute-level multimodal information fusion and helps align the language and vision details more efficiently for fine-grained multimodal features. Extensive experiments show that our method can achieve state-of-the-art performance on ScanRefer and Sr3D/Nr3D datasets.
****
Video-P2P: Video Editing with Cross-attention Control

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8599-8608
Video-P2P is the first framework for real-world video editing with cross-attention control. While attention control has proven effective for image editing with pre-trained image generation models there are currently no large-scale video generation models publicly available. Video-P2P addresses this limitation by adapting an image generation diffusion model to complete various video editing tasks. Specifically we propose to first tune a Text-to-Set (T2S) model to complete an approximate inversion and then optimize a shared unconditional embedding to achieve accurate video inversion with a small memory cost. We further prove that it is crucial for consistent video editing. For attention control we introduce a novel decoupled-guidance strategy which uses different guidance strategies for the source and target prompts. The optimized unconditional embedding for the source prompt improves reconstruction ability while an initialized unconditional embedding for the target prompt enhances editability. Incorporating the attention maps of these two branches enables detailed editing. These technical designs enable various text-driven editing applications including word swap prompt refinement and attention re-weighting. Video-P2P works well on real-world videos for generating new characters while optimally preserving their original poses and scenes. It significantly outperforms previous approaches.
****
Hunting Attributes: Context Prototype-Aware Learning for Weakly Supervised Semantic Segmentation
Feilong Tang, Zhongxing Xu, Zhaojun Qu, Wei Feng, Xingjian Jiang, Zongyuan Ge; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3324-3334
Recent weakly supervised semantic segmentation (WSSS) methods strive to incorporate contextual knowledge to improve the completeness of class activation maps (CAM). In this work we argue that the knowledge bias between instances and contexts affects the capability of the prototype to sufficiently understand instance semantics. Inspired by prototype learning theory we propose leveraging prototype awareness to capture diverse and fine-grained feature attributes of instances. The hypothesis is that contextual prototypes might erroneously activate similar and frequently co-occurring object categories due to this knowledge bias. Therefore we propose to enhance the prototype representation ability by mitigating the bias to better capture spatial coverage in semantic object regions. With this goal we present a Context Prototype-Aware Learning (CPAL) strategy which leverages semantic context to enrich instance comprehension. The core of this method is to accurately capture intra-class variations in object features through context-aware prototypes facilitating the adaptation to the semantic attributes of various instances. We design feature distribution alignment to optimize prototype awareness aligning instance feature distributions with dense features. In addition a unified training framework is proposed to combine label-guided classification supervision and prototypes-guided self-supervision. Experimental results on PASCAL VOC 2012 and MS COCO 2014 show that CPAL significantly improves off-the-shelf methods and achieves state-of-the-art performance. The project is available at \href https://github.com/Barrett-python/CPAL https://github.com/Barrett-python/CPAL.
****
SCINeRF: Neural Radiance Fields from a Snapshot Compressive Image
Yunhao Li, Xiaodong Wang, Ping Wang, Xin Yuan, Peidong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10542-10552
In this paper we explore the potential of Snapshot Com- pressive Imaging (SCI) technique for recovering the under- lying 3D scene representation from a single temporal com- pressed image. SCI is a cost-effective method that enables the recording of high-dimensional data such as hyperspec- tral or temporal information into a single image using low- cost 2D imaging sensors. To achieve this a series of spe- cially designed 2D masks are usually employed which not only reduces sto

rage requirements but also offers potential privacy protection. Inspired by this to take one step further our approach builds upon the powerful 3D scene represe n- tation capabilities of neural radiance fields (NeRF). Specif- ically we formu late the physical imaging process of SCI as part of the training of NeRF allowin g us to exploit its impressive performance in capturing complex scene struc- tur es. To assess the effectiveness of our method we con- duct extensive evaluations using both synthetic data and real data captured by our SCI system. Extensive e xperi- mental results demonstrate that our proposed approach sur- passes the sta te-of-the-art methods in terms of image re- construction and novel view image sy nthesis. Moreover our method also exhibits the ability to restore high frame- ra te multi-view consistent images by leveraging SCI and the rendering capabilities of NeRF. The code is available at https://github.com/WU-CVGL/SCINeRF.
****

PIE-NeRF: Physics-based Interactive Elastodynamics with NeRF

Yutao Feng, Yintong Shang, Xuan Li, Tianjia Shao, Chenfanfu Jiang, Yin Yang; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4450-4461

We show that physics-based simulations can be seamlessly integrated with NeRF to generate high-quality elastodynamics of real-world objects. Unlike existing met hods we discretize nonlinear hyperelasticity in a meshless way obviating the nec essity for intermediate auxiliary shape proxies like a tetrahedral mesh or voxel grid. A quadratic generalized moving least square is employed to capture nonlin ear dynamics and large deformation on the implicit model. Such meshless integrat ion enables versatile simulations of complex and codimensional shapes. We adapti vely place the least-square kernels according to the NeRF density field to signi ficantly reduce the complexity of the nonlinear simulation. As a result physical ly realistic animations can be conveniently synthesized using our method for a w ide range of hyperelastic materials at an interactive rate. For more information please visit https://fytalon.github.io/pienerf.
****

Improved Visual Grounding through Self-Consistent Explanations

Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C. Berg, Vicente Ordon ez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2024, pp. 13095-13105

Vision-and-language models trained to match images with text can be combined wit h visual explanation methods to point to the locations of specific objects in an image. Our work shows that the localization --"grounding'"-- abilities of these models can be further improved by finetuning for self-consistent visual explana tions. We propose a strategy for augmenting existing text-image datasets with pa raphrases using a large language model and SelfEQ a weakly-supervised strategy o n visual explanation maps for paraphrases that encourages self-consistency. Spec ifically for an input textual phrase we attempt to generate a paraphrase and fin etune the model so that the phrase and paraphrase map to the same region in the image. We posit that this both expands the vocabulary that the model is able to handle and improves the quality of the object locations highlighted by gradient-based visual explanation methods (e.g. GradCAM). We demonstrate that SelfEQ impr oves performance on Flickr30k ReferIt and RefCOCO+ over a strong baseline method and several prior works. Particularly comparing to other methods that do not us e any type of box annotations we obtain 84.07% on Flickr30k (an absolute improve ment of 4.69%) 67.40% on ReferIt (an absolute improvement of 7.68%) and 75.10% 5 5.49% on RefCOCO+ test sets A and B respectively (an absolute improvement of 3.7 4% on average).
****

Monkey: Image Resolution and Text Label Are Important Things for Large Multi-mod al Models

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yu liang Liu, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26763-26773

Large Multimodal Models (LMMs) have shown promise in vision-language tasks but s truggle with high-resolution input and detailed scene understanding. Addressing

these challenges we introduce Monkey to enhance LMM capabilities. Firstly Monkey processes input images by dividing them into uniform patches each matching the size (e.g. 448x448) used in the original training of the well-trained vision encoder. Equipped with individual adapter for each patch Monkey can handle higher resolutions up to 1344x896 pixels enabling the detailed capture of complex visual information. Secondly it employs a multi-level description generation method enriching the context for scene-object associations. This two-part strategy ensures more effective learning from generated data: the higher resolution allows for a more detailed capture of visuals which in turn enhances the effectiveness of comprehensive descriptions. Extensive ablative results validate the effectiveness of our designs. Additionally experiments on 18 datasets further demonstrate that Monkey surpasses existing LMMs in many tasks like Image Captioning and various Visual Question Answering formats. Specially in qualitative tests focused on dense text question answering Monkey has exhibited encouraging results compared with GPT4V. Code is available at https://github.com/Yuliang-Liu/Monkey.
****
FlashAvatar: High-fidelity Head Avatar with Efficient Gaussian Embedding
Jun Xiang, Xuan Gao, Yudong Guo, Juyong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1802-1812
We propose FlashAvatar a novel and lightweight 3D animatable avatar representation that could reconstruct a digital avatar from a short monocular video sequence in minutes and render high-fidelity photo-realistic images at 300FPS on a consumer-grade GPU. To achieve this we maintain a uniform 3D Gaussian field embedded in the surface of a parametric face model and learn extra spatial offset to model non-surface regions and subtle facial details. While full use of geometric priors can capture high-frequency facial details and preserve exaggerated expressions proper initialization can help reduce the number of Gaussians thus enabling super-fast rendering speed. Extensive experimental results demonstrate that FlashAvatar outperforms existing works regarding visual quality and personalized details and is almost an order of magnitude faster in rendering speed. Project page: https://ustc3dv.github.io/FlashAvatar/
****
DifFlow3D: Toward Robust Uncertainty-Aware Scene Flow Estimation with Iterative Diffusion-Based Refinement
Jiuming Liu, Guangming Wang, Weicai Ye, Chaokang Jiang, Jinru Han, Zhe Liu, Guofeng Zhang, Dalong Du, Hesheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15109-15119
Scene flow estimation which aims to predict per-point 3D displacements of dynamic scenes is a fundamental task in the computer vision field. However previous works commonly suffer from unreliable correlation caused by locally constrained searching ranges and struggle with accumulated inaccuracy arising from the coarse-to-fine structure. To alleviate these problems we propose a novel uncertainty-aware scene flow estimation network (DifFlow3D) with the diffusion probabilistic model. Iterative diffusion-based refinement is designed to enhance the correlation robustness and resilience to challenging cases e.g. dynamics noisy inputs repetitive patterns etc. To restrain the generation diversity three key flow-related features are leveraged as conditions in our diffusion model. Furthermore we also develop an uncertainty estimation module within diffusion to evaluate the reliability of estimated scene flow. Our DifFlow3D achieves state-of-the-art performance with 24.0% and 29.1% EPE3D reduction respectively on FlyingThings3D and KITTI 2015 datasets. Notably our method achieves an unprecedented millimeter-level accuracy (0.0078m in EPE3D) on the KITTI dataset. Additionally our diffusion-based refinement paradigm can be readily integrated as a plug-and-play module into existing scene flow networks significantly increasing their estimation accuracy. Codes are released at https://github.com/IRMVLab/DifFlow3D.
****
Decompose-and-Compose: A Compositional Approach to Mitigating Spurious Correlation
Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdan Parast, Hamidreza Yaghoubi Araghi, Mahdieh Soleymani Baghshah; Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27662-27671
While standard Empirical Risk Minimization (ERM) training is proven effective fo r image classification on in-distribution data it fails to perform well on out-o f-distribution samples. One of the main sources of distribution shift for image classification is the compositional nature of images. Specifically in addition t o the main object or component(s) determining the label some other image compone nts usually exist which may lead to the shift of input distribution between trai n and test environments. More importantly these components may have spurious cor relations with the label. To address this issue we propose Decompose-and-Compose (DaC) which improves robustness to correlation shift by a compositional approac h based on combining elements of images. Based on our observations models traine d with ERM usually highly attend to either the causal components or the componen ts having a high spurious correlation with the label (especially in datapoints o n which models have a high confidence). In fact according to the amount of spuri ous correlation and the easiness of classification based on the causal or non-ca usal components the model usually attends to one of these more (on samples with high confidence). Following this we first try to identify the causal components of images using class activation maps of models trained with ERM. Afterward we i ntervene on images by combining them and retraining the model on the augmented d ata including the counterfactual ones. This work proposes a group-balancing meth od by intervening on images without requiring group labels or information regard ing the spurious features during training. The method has an overall better wors t group accuracy compared to previous methods with the same amount of supervisio n on the group labels in correlation shift. Our code is available at https://git hub.com/fhn98/DaC.
****
FlashEval: Towards Fast and Accurate Evaluation of Text-to-image Diffusion Gener ative Models
Lin Zhao, Tianchen Zhao, Zinan Lin, Xuefei Ning, Guohao Dai, Huazhong Yang, Yu W ang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2024, pp. 16122-16131
In recent years there has been significant progress in the development of text-t o-image generative models. Evaluating the quality of the generative models is on e essential step in the development process. Unfortunately the evaluation proces s could consume a significant amount of computational resources making the requi red periodic evaluation of model performance (e.g. monitoring training progress) impractical. Therefore we seek to improve the evaluation efficiency by selectin g the representative subset of the text-image dataset. We systematically investi gate the design choices including the selection criteria (textural features or i magebased metrics) and the selection granularity (prompt-level or set-level). We find that the insights from prior work on subset selection for training data do not generalize to this problem and we propose FlashEval an iterative search alg orithm tailored to evaluation data selection. We demonstrate the effectiveness o f FlashEval on ranking diffusion models with various configurations including ar chitectures quantization levels and sampler schedules on COCO and DiffusionDB da tasets. Our searched 50-item subset could achieve comparable evaluation quality to the randomly sampled 500-item subset for COCO annotations on unseen models ac hieving a 10x evaluation speedup. We release the condensed subset of these commo nly used datasets to help facilitate diffusion algorithm design and evaluation a nd open-source FlashEval as a tool for condensing future datasets accessible at https://github.com/thu-nics/FlashEval.
****
ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images
Yiqi Shi, Duo Liu, Liguo Zhang, Ye Tian, Xuezhi Xia, Xiaojing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024 , pp. 3015-3024
This paper presents a novel zero-shot method for jointly denoising and enhancing real-word low-light images. The proposed method is independent of training data and noise distribution. Guided by illumination we integrate denoising and enhan

cing processes seamlessly enabling end-to-end training. Pairs of downsampled ima ges are extracted from a single original low-light image and processed to prelim inarily reduce noise. Based on the smoothness of illumination near-authentic ill umination can be estimated from the denoised low-light image. Specifically the i llumination is constrained by the denoised image's brightness uniformly amplifyi ng pixels to raise overall brightness to normal-light level. We simultaneously r estrict the illumination by scaling each pixel of the denoised image based on it s intensity controlling the enhancement amplitude for different pixels. Applying the illumination to the original low-light image yields an adaptively enhanced reflection. This prevents under-enhancement and localized overexposure. Notably we concatenate the reflection with the illumination preserving their computation al relationship to ultimately remove noise from the original low-light image in the form of reflection. This provides sufficient image information for the denoi sing procedure without changing the noise characteristics. Extensive experiments demonstrate that our method outperforms other state-of-the-art methods. The sou rce code is available at https://github.com/Doyle59217/ZeroIG.
****

View From Above: Orthogonal-View aware Cross-view Localization
Shan Wang, Chuong Nguyen, Jiawei Liu, Yanhao Zhang, Sundaram Muthu, Fahira Afzal Maken, Kaihao Zhang, Hongdong Li; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2024, pp. 14843-14852
This paper presents a novel aerial-to-ground feature aggregation strategy tailor ed for the task of cross-view image-based geo-localization. Conventional vision-based methods heavily rely on matching ground-view image features with a pre-rec orded image database often through establishing planar homography correspondence s via a planar ground assumption. As such they tend to ignore features that are off-ground and not suited for handling visual occlusions leading to unreliable l ocalization in challenging scenarios. We propose a Top-to-Ground Aggregation mod ule that capitalizes aerial orthographic views to aggregate features down to the ground level leveraging reliable off-ground information to improve feature alig nment. Furthermore we introduce a Cycle Domain Adaptation loss that ensures feat ure extraction robustness across domain changes. Additionally an Equidistant Re-projection loss is introduced to equalize the impact of all keypoints on orienta tion error leading to a more extended distribution of keypoints which benefits o rientation estimation. On both KITTI and Ford Multi-AV datasets our method consi stently achieves the lowest mean longitudinal and lateral translations across di fferent settings and obtains the smallest orientation error when the initial pos e is less accurate a more challenging setting. Further it can complete an entire route through continual vehicle pose estimation with initial vehicle pose given only at the starting point.
****

FinePOSE: Fine-Grained Prompt-Driven 3D Human Pose Estimation via Diffusion Mode ls
Jinglin Xu, Yijie Guo, Yuxin Peng; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2024, pp. 561-570
The 3D Human Pose Estimation (3D HPE) task uses 2D images or videos to predict h uman joint coordinates in 3D space. Despite recent advancements in deep learning -based methods they mostly ignore the capability of coupling accessible texts an d naturally feasible knowledge of humans missing out on valuable implicit superv ision to guide the 3D HPE task. Moreover previous efforts often study this task from the perspective of the whole human body neglecting fine-grained guidance hi dden in different body parts. To this end we present a new Fine-Grained Prompt-D riven Denoiser based on a diffusion model for 3D HPE named FinePOSE. It consists of three core blocks enhancing the reverse process of the diffusion model: (1) Fine-grained Part-aware Prompt learning (FPP) block constructs fine-grained part -aware prompts via coupling accessible texts and naturally feasible knowledge of body parts with learnable prompts to model implicit guidance. (2) Fine-grained Prompt-pose Communication (FPC) block establishes fine-grained communications be tween learned part-aware prompts and poses to improve the denoising quality. (3) Prompt-driven Timestamp Stylization (PTS) block integrates learned prompt embed

ding and temporal information related to the noise level to enable adaptive adju stment at each denoising step. Extensive experiments on public single-human pose estimation datasets show that FinePOSE outperforms state-of-the-art methods. We further extend FinePOSE to multi-human pose estimation. Achieving 34.3mm averag e MPJPE on the EgoHumans dataset demonstrates the potential of FinePOSE to deal with complex multi-human scenarios. Code is available at https://github.com/PKU-ICST-MIPL/FinePOSE_CVPR2024.
****
BEM: Balanced and Entropy-based Mix for Long-Tailed Semi-Supervised Learning
Hongwei Zheng, Linyuan Zhou, Han Li, Jinming Su, Xiaoming Wei, Xiaoming Xu; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2024, pp. 22893-22903
Data mixing methods play a crucial role in semi-supervised learning (SSL) but th eir application is unexplored in long-tailed semi-supervised learning (LTSSL). T he primary reason is that the in-batch mixing manner fails to address class imba lance. Furthermore existing LTSSL methods mainly focus on re-balancing data quan tity but ignore class-wise uncertainty which is also vital for class balance. Fo r instance some classes with sufficient samples might still exhibit high uncerta inty due to indistinguishable features. To this end this paper introduces the Ba lanced and Entropy-based Mix (BEM) a pioneering mixing approach to re-balance th e class distribution of both data quantity and uncertainty. Specifically we firs t propose a class balanced mix bank to store data of each class for mixing. This bank samples data based on the estimated quantity distribution thus re-balancin g data quantity. Then we present an entropy-based learning approach to re-balanc e class-wise uncertainty including entropy-based sampling strategy entropy-based selection module and entropy-based class balanced loss. Our BEM first leverages data mixing for improving LTSSL and it can also serve as a complement to the ex isting re-balancing methods. Experimental results show that BEM significantly en hances various LTSSL frameworks and achieves state-of-the-art performances acros s multiple benchmarks.
****
HUGS: Holistic Urban 3D Scene Understanding via Gaussian Splatting
Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wa ng, Andreas Geiger, Yiyi Liao; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2024, pp. 21336-21345
Holistic understanding of urban scenes based on RGB images is a challenging yet important problem. It encompasses understanding both the geometry and appearance to enable novel view synthesis parsing semantic labels and tracking moving obje cts. Despite considerable progress existing approaches often focus on specific a spects of this task and require additional inputs such as LiDAR scans or manuall y annotated 3D bounding boxes. In this paper we introduce a novel pipeline that utilizes 3D Gaussian Splatting for holistic urban scene understanding. Our main idea involves the joint optimization of geometry appearance semantics and motion using a combination of static and dynamic 3D Gaussians where moving object pose s are regularized via physical constraints. Our approach offers the ability to r ender new viewpoints in real-time yielding 2D and 3D semantic information with h igh accuracy and reconstruct dynamic scenes even in scenarios where 3D bounding box detection are highly noisy. Experimental results on KITTI KITTI-360 and Virt ual KITTI 2 demonstrate the effectiveness of our approach. Our project page is a t https://xdimlab.github.io/hugs_website.
****
DreamPropeller: Supercharge Text-to-3D Generation with Parallel Sampling
Linqi Zhou, Andy Shih, Chenlin Meng, Stefano Ermon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4610-461 9
Recent methods such as Score Distillation Sampling (SDS) and Variational Score D istillation (VSD) using 2D diffusion models for text-to-3D generation have demon strated impressive generation quality. However the long generation time of such algorithms significantly degrades the user experience. To tackle this problem we propose DreamPropeller a drop-in acceleration algorithm that can be wrapped aro

und any existing text-to-3D generation pipeline based on score distillation. Our framework generalizes Picard iterations a classical algorithm for parallel sampling an ODE path and can account for non-ODE paths such as momentum-based gradient updates and changes in dimensions during the optimization process as in many cases of 3D generation. We show that our algorithm trades parallel compute for wallclock time and empirically achieves up to 4.7x speedup with a negligible drop in generation quality for all tested frameworks.
****
PeVL: Pose-Enhanced Vision-Language Model for Fine-Grained Human Action Recognition
Haosong Zhang, Mei Chee Leong, Liyuan Li, Weisi Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18857-18867
Recent progress in Vision-Language (VL) foundation models has revealed the great advantages of cross-modality learning. However due to a large gap between vision and text they might not be able to sufficiently utilize the benefits of cross-modality information. In the field of human action recognition the additional pose modality may bridge the gap between vision and text to improve the effectiveness of cross-modality learning. In this paper we propose a novel framework called the Pose-enhanced Vision-Language (PeVL) model to adapt the VL model with pose modality to learn effective knowledge of fine-grained human actions. Our PeVL model includes two novel components: an Unsymmetrical Cross-Modality Refinement (UCMR) block and a Semantic-Guided Multi-level Contrastive (SGMC) module. The UCMR block includes Pose-guided Visual Refinement (P2V-R) and Visual-enriched Pose Refinement (V2P-R) for effective cross-modality learning. The SGMC module includes Multi-level Contrastive Associations of vision-text and pose-text at both action and sub-action levels and a Semantic-Guided Loss enabling effective contrastive learning with text. Built upon a pre-trained VL foundation model our model integrates trainable adapters and can be trained end-to-end. Our novel PeVL design over VL foundation model yields remarkable performance gains on four fine- grained human action recognition datasets achieving a new SOTA with a significantly small number of FLOPs for low- cost re-training.
****
DeepCache: Accelerating Diffusion Models for Free
Xinyin Ma, Gongfan Fang, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15762-15772
Diffusion models have recently gained unprecedented attention in the field of image synthesis due to their remarkable generative capabilities. Notwithstanding their prowess these models often incur substantial computational costs primarily attributed to the sequential denoising process and cumbersome model size. Traditional methods for compressing diffusion models typically involve extensive retraining presenting cost and feasibility challenges. In this paper we introduce DeepCache a novel training-free paradigm that accelerates diffusion models from the perspective of model architecture. DeepCache capitalizes on the inherent temporal redundancy observed in the sequential denoising steps of diffusion models which caches and retrieves features across adjacent denoising stages thereby curtailing redundant computations. Utilizing the property of the U-Net we reuse the high-level features while updating the low-level features in a very cheap way. This innovative strategy in turn enables a speedup factor of 2.3xfor Stable Diffusion v1.5 with only a 0.05 decline in CLIP Score and 4.1xfor LDM-4-G with a slight decrease of 0.22 in FID on ImageNet. Our experiments also demonstrate DeepCache's superiority over existing pruning and distillation methods that necessitate retraining and its compatibility with current sampling techniques. Furthermore we find that under the same throughput DeepCache effectively achieves comparable or even marginally improved results with DDIM or PLMS.
****
GeoAuxNet: Towards Universal 3D Representation Learning for Multi-sensor Point Clouds
Shengjun Zhang, Xin Fei, Yueqi Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20019-20028

Point clouds captured by different sensors such as RGB-D cameras and LiDAR possess non-negligible domain gaps. Most existing methods design different network architectures and train separately on point clouds from various sensors. Typically point-based methods achieve outstanding performances on even-distributed dense point clouds from RGB-D cameras while voxel-based methods are more efficient for large-range sparse LiDAR point clouds. In this paper we propose geometry-to-voxel auxiliary learning to enable voxel representations to access point-level geometric information which supports better generalisation of the voxel-based backbone with additional interpretations of multi-sensor point clouds. Specifically we construct hierarchical geometry pools generated by a voxel-guided dynamic point network which efficiently provide auxiliary fine-grained geometric information adapted to different stages of voxel features. We conduct experiments on joint multi-sensor datasets to demonstrate the effectiveness of GeoAuxNet. Enjoying elaborate geometric information our method outperforms other models collectively trained on multi-sensor datasets and achieve competitive results with the-state-of-art experts on each single dataset.
****
Unveiling the Power of Audio-Visual Early Fusion Transformers with Dense Interactions through Masked Modeling
Shentong Mo, Pedro Morgado; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27186-27196
Humans possess a remarkable ability to integrate auditory and visual information enabling a deeper understanding of the surrounding environment. This early fusion of audio and visual cues demonstrated through cognitive psychology and neuroscience research offers promising potential for developing multimodal perception models. However training early fusion architectures poses significant challenges as the increased model expressivity requires robust learning frameworks to harness their enhanced capabilities. In this paper we address this challenge by leveraging the masked reconstruction framework previously successful in unimodal settings to train audio-visual encoders with early fusion. Additionally we propose an attention-based fusion module that captures interactions between local audio and visual representations enhancing the model's ability to capture fine-grained interactions. While effective this procedure can become computationally intractable as the number of local representations increases. Thus to address the computational complexity we propose an alternative procedure that factorizes the local representations before representing audio-visual interactions. Extensive evaluations on a variety of datasets demonstrate the superiority of our approach in audio-event classification visual sound localization sound separation and audio-visual segmentation. These contributions enable the efficient training of deeply integrated audio-visual models and significantly advance the usefulness of early fusion architectures.
****
Learning Correlation Structures for Vision Transformers
Manjin Kim, Paul Hongsuck Seo, Cordelia Schmid, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18941-18951
We introduce a new attention mechanism dubbed structural self-attention (StructSA) that leverages rich correlation patterns naturally emerging in key-query interactions of attention. StructSA generates attention maps by recognizing space-time structures of key-query correlations via convolution and uses them to dynamically aggregate local contexts of value features. This effectively leverages rich structural patterns in images and videos such as scene layouts object motion and inter-object relations.Using StructSA as a main building block we develop the structural vision transformer (StructViT) and evaluate its effectiveness on both image and video classification tasks achieving state-of-the-art results on ImageNet-1K Kinetics-400 Something-Something V1 & V2 Diving-48 and FineGym.
****
Dysen-VDM: Empowering Dynamics-aware Text-to-Video Diffusion with LLMs
Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Tat-Seng Chua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp

. 7641-7653
Text-to-video (T2V) synthesis has gained increasing attention in the community i
n which the recently emerged diffusion models (DMs) have promisingly shown stron
ger performance than the past approaches. While existing state-of-the-art DMs ar
e competent to achieve high-resolution video generation they may largely suffer
from key limitations (e.g. action occurrence disorders crude video motions) with
 respect to the intricate temporal dynamics modeling one of the crux of video sy
nthesis. In this work we investigate strengthening the awareness of video dynami
cs for DMs for high-quality T2V generation. Inspired by human intuition we desig
n an innovative dynamic scene manager (dubbed as Dysen) module which includes (s
tep-1) extracting from input text the key actions with proper time-order arrange
ment (step-2) transforming the action schedules into the dynamic scene graph (DS
G) representations and (step-3) enriching the scenes in the DSG with sufficient
and reasonable details. Taking advantage of the existing powerful LLMs (e.g. Cha
tGPT) via in-context learning Dysen realizes (nearly) human-level temporal dynam
ics understanding. Finally the resulting video DSG with rich action scene detail
s is encoded as fine-grained spatio-temporal features integrated into the backbo
ne T2V DM for video generating. Experiments on popular T2V datasets suggest that
 our Dysen-VDM consistently outperforms prior arts with significant margins espe
cially in scenarios with complex actions.
****
PrPSeg: Universal Proposition Learning for Panoramic Renal Pathology Segmentatio
n
Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Jialin Yue, Juming Xiong, Lining
Yu, Yifei Wu, Mengmeng Yin, Yu Wang, Shilin Zhao, Yucheng Tang, Haichun Yang, Yu
ankai Huo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2024, pp. 11736-11746
Understanding the anatomy of renal pathology is crucial for advancing disease di
agnostics treatment evaluation and clinical research. The complex kidney system
comprises various components across multiple levels including regions (cortex me
dulla) functional units (glomeruli tubules) and cells (podocytes mesangial cells
 in glomerulus). Prior studies have predominantly overlooked the intricate spati
al interrelations among objects from clinical knowledge. In this research we int
roduce a novel universal proposition learning approach called panoramic renal pa
thology segmentation (PrPSeg) designed to segment comprehensively panoramic stru
ctures within kidney by integrating extensive knowledge of kidney anatomy. In th
is paper we propose (1) the design of a comprehensive universal proposition matr
ix for renal pathology facilitating the incorporation of classification and spat
ial relationships into the segmentation process; (2) a token-based dynamic head
single network architecture with the improvement of the partial label image segm
entation and capability for future data enlargement; and (3) an anatomy loss fun
ction quantifying the inter-object relationships across the kidney.
****
RepKPU: Point Cloud Upsampling with Kernel Point Representation and Deformation
Yi Rong, Haoran Zhou, Kang Xia, Cheng Mei, Jiahao Wang, Tong Lu; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 21050-21060
In this work we present RepKPU an efficient network for point cloud upsampling.
We propose to promote upsampling performance by exploiting better shape represen
tation and point generation strategy. Inspired by KPConv we propose a novel repr
esentation called RepKPoints to effectively characterize the local geometry whos
e advantages over prior representations are as follows: (1) density-sensitive; (
2) large receptive fields; (3) position-adaptive which makes RepKPoints a genera
lized form of previous representations. Moreover we propose a novel paradigm nam
ely Kernel-to-Displacement generation for point generation where point cloud ups
ampling is reformulated as the deformation of kernel points. Specifically we pro
pose KP-Queries which is a set of kernel points with predefined positions and le
arned features to serve as the initial state of upsampling. Using cross-attentio
n mechanisms we achieve interactions between RepKPoints and KP-Queries and subse
quently KP-Queries are converted to displacement features followed by a MLP to p

redict the new positions of KP-Queries which serve as the generated points. Exte
nsive experimental results demonstrate that RepKPU outperforms state-of-the-art
methods on several widely-used benchmark datasets with high efficiency.
****
ConCon-Chi: Concept-Context Chimera Benchmark for Personalized Vision-Language T
asks
Andrea Rosasco, Stefano Berti, Giulia Pasquale, Damiano Malafronte, Shogo Sato,
Hiroyuki Segawa, Tetsugo Inada, Lorenzo Natale; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22239-22248
While recent Vision-Language (VL) models excel at open-vocabulary tasks it is un
clear how to use them with specific or uncommon concepts. Personalized Text-to-I
mage Retrieval (TIR) or Generation (TIG) are recently introduced tasks that repr
esent this challenge where the VL model has to learn a concept from few images a
nd respectively discriminate or generate images of the target concept in arbitra
ry contexts. We identify the ability to learn new meanings and their composition
ality with known ones as two key properties of a personalized system. We show th
at the available benchmarks offer a limited validation of personalized textual c
oncept learning from images with respect to the above properties and introduce C
onCon-Chi as a benchmark for both personalized TIR and TIG designed to fill this
 gap. We modelled the new-meaning concepts by crafting chimeric objects and form
ulating a large varied set of contexts where we photographed each object. To pro
mote the compositionality assessment of the learned concepts with known contexts
 we combined different contexts with the same concept and vice-versa. We carry o
ut a thorough evaluation of state-of-the-art methods on the resulting dataset. O
ur study suggests that future work on personalized TIR and TIG methods should fo
cus on the above key properties and we propose principles and a dataset for thei
r performance assessment. Dataset: https://doi.org/10.48557/QJ1166 and code: htt
ps://github.com/hsp-iit/concon-chi_benchmark.
****
Weakly-Supervised Audio-Visual Video Parsing with Prototype-based Pseudo-Labelin
g
Kranthi Kumar Rachavarapu, Kalyan Ramakrishnan, Rajagopalan A. N.; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 18952-18962
In this paper we address the weakly-supervised Audio-Visual Video Parsing (AVVP)
 problem which aims at labeling events in a video as audible visible or both and
 temporally localizing and classifying them into known categories. This is chall
enging since we only have access to video-level (weak) event labels when trainin
g but need to predict event labels at the segment (frame) level at test time. Re
cent methods employ multiple-instance learning (MIL) techniques that tend to foc
us solely on the most discriminative segments resulting in frequent misclassific
ations. Our idea is to first construct several prototype features for each event
 class by clustering key segments identified for the event in the training data.
 We then assign pseudo labels to all training segments based on their feature si
milarities with these prototypes and re-train the model under weak and strong su
pervision. We facilitate this by structuring the feature space with contrastive
learning using pseudo labels. Experiments show that we outperform existing metho
ds for weakly-supervised AVVP. We also show that learning with weak and iterativ
ely re-estimated pseudo labels can be interpreted as an expectation-maximization
 (EM) algorithm providing further insight for our training procedure.
****
Intraoperative 2D/3D Image Registration via Differentiable X-ray Rendering
Vivek Gopalakrishnan, Neel Dey, Polina Golland; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11662-11672
Surgical decisions are informed by aligning rapid portable 2D intraoperative ima
ges (e.g. X-rays) to a high-fidelity 3D preoperative reference scan (e.g. CT). H
owever 2D/3D registration can often fail in practice: conventional optimization
methods are prohibitively slow and susceptible to local minima while neural netw
orks trained on small datasets fail on new patients or require impractical landm
ark supervision. We present DiffPose a self-supervised approach that leverages p

atient-specific simulation and differentiable physics-based rendering to achieve accurate 2D/3D registration without relying on manually labeled data. Preoperatively a CNN is trained to regress the pose of a randomly oriented synthetic X-ray rendered from the preoperative CT. The CNN then initializes rapid intraoperative test-time optimization that uses the differentiable X-ray renderer to refine the solution. Our work further proposes several geometrically principled methods for sampling camera poses from SE(3) for sparse differentiable rendering and for driving registration in the tangent space se(3) with geodesic and multiscale locality-sensitive losses. DiffPose achieves sub-millimeter accuracy across surgical datasets at intraoperative speeds improving upon existing unsupervised methods by an order of magnitude and even outperforming supervised baselines. Our implementation is at https://github.com/eigenvivek/DiffPose.
****
MICap: A Unified Model for Identity-Aware Movie Descriptions
Haran Raajesh, Naveen Reddy Desanur, Zeeshan Khan, Makarand Tapaswi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14011-14021
Characters are an important aspect of any storyline and identifying and including them in descriptions is necessary for story understanding. While previous work has largely ignored identity and generated captions with someone (anonymized names) recent work formulates id-aware captioning as a fill-in-the-blanks (FITB) task where given a caption with blanks the goal is to predict person id labels. However to predict captions with ids a two-stage approach is required: first predict captions with someone then fill in identities. In this work we present a new single stage approach that can seamlessly switch between id-aware caption generation or FITB when given a caption with blanks. Our model Movie-Identity Captioner (MICap) uses a shared auto-regressive decoder that benefits from training with FITB and full-caption generation objectives while the encoder can benefit from or disregard captions with blanks as input. Another challenge with id-aware captioning is the lack of a metric to capture subtle differences between person ids. To this end we introduce iSPICE a caption evaluation metric that focuses on identity tuples created through intermediate scene graphs. We evaluate MICap on Large-Scale Movie Description Challenge (LSMDC) where we show a 4.2% improvement in FITB accuracy and a 1-2% bump in classic captioning metrics.
****
MonoDiff: Monocular 3D Object Detection and Pose Estimation with Diffusion Models
Yasiru Ranasinghe, Deepti Hegde, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10659-10670
3D object detection and pose estimation from a single-view image is challenging due to the high uncertainty caused by the absence of 3D perception. As a solution recent monocular 3D detection methods leverage additional modalities such as stereo image pairs and LiDAR point clouds to enhance image features at the expense of additional annotation costs. We propose using diffusion models to learn effective representations for monocular 3D detection without additional modalities or training data. We present MonoDiff a novel framework that employs the reverse diffusion process to estimate 3D bounding box and orientation. But considering the variability in bounding box sizes along different dimensions it is ineffective to sample noise from a standard Gaussian distribution. Hence we adopt a Gaussian mixture model to sample noise during the forward diffusion process and initialize the reverse diffusion process. Furthermore since the diffusion model generates the 3D parameters for a given object image we leverage 2D detection information to provide additional supervision by maintaining the correspondence between 3D/2D projection. Finally depending on the signal-to-noise ratio we incorporate a dynamic weighting scheme to account for the level of uncertainty in the supervision by projection at different timesteps. MonoDiff outperforms current state-of-the-art monocular 3D detection methods on the KITTI and Waymo benchmarks without additional depth priors. MonoDiff project is available at: https://dylran.github.io/monodiff.github.io.

****
General Object Foundation Model for Images and Videos at Scale
Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, Song Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3783-3795
We present GLEE in this work an object-level foundation model for locating and identifying objects in images and videos. Through a unified framework GLEEaccomplishes detection segmentation tracking grounding and identification of arbitrary objects in the open world scenario for various object perception tasks. Adopting a cohesive learning strategy GLEE acquires knowledge from diverse data sources with varying supervision levels to formulate general object representations excelling in zero-shot transfer to new data and tasks. Specifically we employ an image encoder text encoder and visual prompter to handle multi-modal inputs enabling to simultaneously solve various object-centric downstream tasks while maintaining state-of-the-art performance. Demonstrated through extensive training on over five million images from diverse benchmarks GLEE exhibits remarkable versatility and improved generalization performance efficiently tackling downstream tasks without the need for task-specific adaptation. By integrating large volumes of automatically labeled data we further enhance its zero-shot generalization capabilities. Additionally GLEE is capable of being integrated into Large Language Models serving as a foundational model to provide universal object-level information for multi-modal tasks. We hope that the versatility and universality of our method will mark a significant step in the development of efficient visual foundation models for AGI systems. The models and code are released at https://github.com/FoundationVision/GLEE.
****
An Upload-Efficient Scheme for Transferring Knowledge From a Server-Side Pre-trained Generator to Clients in Heterogeneous Federated Learning
Jianqing Zhang, Yang Liu, Yang Hua, Jian Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12109-12119
Heterogeneous Federated Learning (HtFL) enables collaborative learning on multiple clients with different model architectures while preserving privacy. Despite recent research progress knowledge sharing in HtFL is still difficult due to data and model heterogeneity. To tackle this issue we leverage the knowledge stored in pre-trained generators and propose a new upload-efficient knowledge transfer scheme called Federated Knowledge-Transfer Loop (FedKTL). Our FedKTL can produce client-task-related prototypical image-vector pairs via the generator's inference on the server. With these pairs each client can transfer pre-existing knowledge from the generator to its local model through an additional supervised local task. We conduct extensive experiments on four datasets under two types of data heterogeneity with 14 kinds of models including CNNs and ViTs. Results show that our upload-efficient FedKTL surpasses seven state-of-the-art methods by up to 7.31% in accuracy. Moreover our knowledge transfer scheme is applicable in scenarios with only one edge client. Code: https://github.com/TsingZ0/FedKTL
****
MeshGPT: Generating Triangle Meshes with Decoder-Only Transformers
Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19615-19625
We introduce MeshGPT a new approach for generating triangle meshes that reflects the compactness typical of artist-created meshes in contrast to dense triangle meshes extracted by iso-surfacing methods from neural fields. Inspired by recent advances in powerful large language models we adopt a sequence-based approach to autoregressively generate triangle meshes as sequences of triangles. We first learn a vocabulary of latent quantized embeddings using graph convolutions which inform these embeddings of the local mesh geometry and topology. These embeddings are sequenced and decoded into triangles by a decoder ensuring that they can effectively reconstruct the mesh. A transformer is then trained on this learned vocabulary to predict the index of the next embedding given previous embeddings.

Once trained our model can be autoregressively sampled to generate new triangle meshes directly generating compact meshes with sharp edges more closely imitating the efficient triangulation patterns of human-crafted meshes. MeshGPT demonstrates a notable improvement over state of the art mesh generation methods with a 9% increase in shape coverage and a 30-point enhancement in FID scores across various categories.
****
Inlier Confidence Calibration for Point Cloud Registration
Yongzhe Yuan, Yue Wu, Xiaolong Fan, Maoguo Gong, Qiguang Miao, Wenping Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5312-5321
Inliers estimation constitutes a pivotal step in partially overlapping point cloud registration. Existing methods broadly obey coordinate-based scheme where inlier confidence is scored through simply capturing coordinate differences in the context. However this scheme results in massive inlier misinterpretation readily consequently affecting the registration performance. In this paper we explore to extend a new definition called inlier confidence calibration (ICC) to alleviate the above issues. Firstly we provide finely initial correspondences for ICC in order to generate high quality reference point cloud copy corresponding to the source point cloud. In particular we develop a soft assignment matrix optimization theorem that offers faster speed and greater precision compared to Sinkhorn. Benefiting from the high quality reference copy we argue the neighborhood patch formed by inlier and its neighborhood should have consistency between source point cloud and its reference copy. Based on this insight we construct transformation-invariant geometric constraints and capture geometric structure consistency to calibrate inlier confidence for estimated correspondences between source point cloud and its reference copy. Finally transformation is further calculated by the weighted SVD algorithm with the calibrated inlier confidence. Our model is trained in an unsupervised manner and extensive experiments on synthetic and real-world datasets illustrate the effectiveness of the proposed method.
****
Instance-aware Exploration-Verification-Exploitation for Instance ImageGoal Navigation
Xiaohan Lei, Min Wang, Wengang Zhou, Li Li, Houqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16329-16339
As a new embodied vision task Instance ImageGoal Navigation (IIN) aims to navigate to a specified object depicted by a goal image in an unexplored environment. The main challenge of this task lies in identifying the target object from different viewpoints while rejecting similar distractors. Existing ImageGoal Navigation methods usually adopt the simple Exploration-Exploitation framework and ignore the identification of specific instance during navigation. In this work we propose to imitate the human behaviour of "getting closer to confirm" when distinguishing objects from a distance. Specifically we design a new modular navigation framework named Instance-aware Exploration-Verification-Exploitation (IEVE) for instancelevel image goal navigation. Our method allows for active switching among the exploration verification and exploitation actions thereby facilitating the agent in making reasonable decisions under different situations. On the challenging HabitatMatterport 3D semantic (HM3DSEM) dataset our method surpasses previous state-of-theart work with a classical segmentation model (0.684 vs. 0.561 success) or a robust model (0.702 vs. 0.561 success). Our code will be made publicly available at https://github.com/XiaohanLei/IEVE.
****
One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion
Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10072-10083
Recent advancements in open-world 3D object generation have been remarkable with image-to-3D methods offering superior fine-grained control over their text-to-3

D counterparts. However most existing models fall short in simultaneously provid ing rapid generation speeds and high fidelity to input images - two features ess ential for practical applications. In this paper we present One-2-3-45++ an inno vative method that transforms a single image into a detailed 3D textured mesh in approximately one minute. Our approach aims to fully harness the extensive know ledge embedded in 2D diffusion models and priors from valuable yet limited 3D da ta. This is achieved by initially finetuning a 2D diffusion model for consistent multi-view image generation followed by elevating these images to 3D with the a id of multi-view-conditioned 3D native diffusion models. Extensive experimental evaluations demonstrate that our method can produce high-quality diverse 3D asse ts that closely mirror the original input image.
****
Image Restoration by Denoising Diffusion Models with Iteratively Preconditioned Guidance
Tomer Garber, Tom Tirer; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2024, pp. 25245-25254
Training deep neural networks has become a common approach for addressing image restoration problems. An alternative for training a "task-specific" network for each observation model is to use pretrained deep denoisers for imposing only the signal's prior within iterative algorithms without additional training. Recentl y a sampling-based variant of this approach has become popular with the rise of diffusion/score-based generative models. Using denoisers for general purpose res toration requires guiding the iterations to ensure agreement of the signal with the observations. In low-noise settings guidance that is based on back-projectio n (BP) has been shown to be a promising strategy (used recently also under the n ames "pseudoinverse" or "range/-space" guidance). However the presence of noise in the observations hinders the gains from this approach. In this paper we propo se a novel guidance technique based on preconditioning that allows traversing fr om BP-based guidance to least squares based guidance along the restoration schem e. The proposed approach is robust to noise while still having much simpler impl ementation than alternative methods (e.g. it does not require SVD or a large num ber of iterations). We use it within both an optimization scheme and a sampling-based scheme and demonstrate its advantages over existing methods for image debl urring and super-resolution.
****
Let's Think Outside the Box: Exploring Leap-of-Thought in Large Language Models with Creative Humor Generation
Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zi tnik, Pan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2024, pp. 13246-13257
Chain-of-Thought (CoT) guides large language models (LLMs) to reason step-by-ste p and can motivate their logical reasoning ability. While effective for logical tasks CoT is not conducive to creative problem-solving which often requires out-of-box thoughts and is crucial for innovation advancements. In this paper we exp lore the Leap-of-Thought (LoT) abilities within LLMs -- a non-sequential creativ e paradigm involving strong associations and knowledge leaps. To this end we stu dy LLMs on the popular Oogiri game which needs participants to have good creativ ity and strong associative thinking for responding unexpectedly and humorously t o the given image text or both and thus is suitable for LoT study. Then to inves tigate LLMs' LoT ability in the Oogiri game we first build a multimodal and mult ilingual Oogiri-GO dataset which contains over 130000 samples from the Oogiri ga me and observe the insufficient LoT ability or failures of most existing LLMs on the Oogiri game. Accordingly we introduce a creative Leap-of-Thought (CLoT) par adigm to improve LLM's LoT ability. CLoT first formulates the Oogiri-GO dataset into LoT-oriented instruction tuning data to train pretrained LLM for achieving certain LoT humor generation and discrimination abilities. Then CLoT designs an explorative self-refinement that encourages the LLM to generate more creative Lo T data via exploring parallels between seemingly unrelated concepts and selects high-quality data to train itself for self-refinement. CLoT not only excels in h umor generation in the Oogiri game as shown in Fig. 1 but also boosts creative a

bilities in various tasks like "cloud guessing game" and "divergent association task". These findings advance our understanding and offer a pathway to improve LLMs' creative capacities for innovative applications across domains. The dataset code and models have been released online: https://zhongshsh.github.io/CLoT.
****

SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes
Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, Francis Engelmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14531-14542
Existing 3D scene understanding methods are heavily focused on 3D semantic and instance segmentation. However identifying objects and their parts only constitutes an intermediate step towards a more fine-grained goal which is effectively interacting with the functional interactive elements (e.g. handles knobs buttons) in the scene to accomplish diverse tasks. To this end we introduce SceneFun3D a large-scale dataset with more than 14.8k highly accurate interaction annotations for 710 high-resolution real-world 3D indoor scenes. We accompany the annotations with motion parameter information describing how to interact with these elements and a diverse set of natural language descriptions of tasks that involve manipulating them in the scene context. To showcase the value of our dataset we introduce three novel tasks namely functionality segmentation task-driven affordance grounding and 3D motion estimation and adapt existing state-of-the-art methods to tackle them. Our experiments show that solving these tasks in real 3D scenes remains challenging despite recent progress in closed-set and open-set 3D scene understanding methods.
****

Readout Guidance: Learning Control from Diffusion Features
Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, Aleksander Holynski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8217-8227
We present Readout Guidance a method for controlling text-to-image diffusion models with learned signals. Readout Guidance uses readout heads lightweight networks trained to extract signals from the features of a pre-trained frozen diffusion model at every timestep. These readouts can encode single-image properties such as pose depth and edges; or higher-order properties that relate multiple images such as correspondence and appearance similarity. Furthermore by comparing the readout estimates to a user-defined target and back-propagating the gradient through the readout head these estimates can be used to guide the sampling process. Compared to prior methods for conditional generation Readout Guidance requires significantly fewer added parameters and training samples and offers a convenient and simple recipe for reproducing different forms of conditional control under a single framework with a single architecture and sampling procedure. We showcase these benefits in the applications of drag-based manipulation identity-consistent generation and spatially aligned control.
****

A Unified Approach for Text- and Image-guided 4D Scene Generation
Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, Shalini De Mello; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7300-7309
Large-scale diffusion generative models are greatly simplifying image video and 3D asset creation from user provided text prompts and images. However the challenging problem of text-to-4D dynamic 3D scene generation with diffusion guidance remains largely unexplored. We propose Dream-in-4D which features a novel two-stage approach for text-to-4D synthesis leveraging (1) 3D and 2D diffusion guidance to effectively learn a high-quality static 3D asset in the first stage; (2) a deformable neural radiance field that explicitly disentangles the learned static asset from its deformation preserving quality during motion learning; and (3) a multi-resolution feature grid for the deformation field with a displacement total variation loss to effectively learn motion with video diffusion guidance in the second stage. Through a user preference study we demonstrate that our approach significantly advances image and motion quality 3D consistency and text fideli

ty for text-to-4D generation compared to baseline approaches. Thanks to its moti
on-disentangled representation Dream-in-4D can also be easily adapted for contro
llable generation where appearance is defined by one or multiple images without
the need to modify the motion learning stage. Thus our method offers for the fir
st time a unified approach for text-to-4D image-to-4D and personalized 4D genera
tion tasks.
****
GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via
Animatable 3D Gaussians
Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zh
ang, Liqiang Nie; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2024, pp. 634-644
We present GaussianAvatar an efficient approach to creating realistic human avat
ars with dynamic 3D appearances from a single video. We start by introducing ani
matable 3D Gaussians to explicitly represent humans in various poses and clothin
g styles. Such an explicit and animatable representation can fuse 3D appearances
 more efficiently and consistently from 2D observations. Our representation is f
urther augmented with dynamic properties to support pose-dependent appearance mo
deling where a dynamic appearance network along with an optimizable feature tens
or is designed to learn the motion-to-appearance mapping. Moreover by leveraging
 the differentiable motion condition our method enables a joint optimization of
motions and appearances during avatar modeling which helps to tackle the long-st
anding issue of inaccurate motion estimation in monocular settings. The efficacy
 of GaussianAvatar is validated on both the public dataset and our collected dat
aset demonstrating its superior performances in terms of appearance quality and
rendering efficiency.
****
MTMMC: A Large-Scale Real-World Multi-Modal Camera Tracking Benchmark
Sanghyun Woo, Kwanyong Park, Inkyu Shin, Myungchul Kim, In So Kweon; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 22335-22346
Multi-target multi-camera tracking is a crucial task that involves identifying a
nd tracking individuals over time using video streams from multiple cameras. Thi
s task has practical applications in various fields such as visual surveillance
crowd behavior analysis and anomaly detection. However due to the difficulty and
 cost of collecting and labeling data existing datasets for this task are either
 synthetically generated or artificially constructed within a controlled camera
network setting which limits their ability to model real-world dynamics and gene
ralize to diverse camera configurations. To address this issue we present MTMMC
a real-world large-scale dataset that includes long video sequences captured by
16 multi-modal cameras in two different environments - campus and factory - acro
ss various time weather and season conditions. This dataset provides a challengi
ng test bed for studying multi-camera tracking under diverse real-world complexi
ties and includes an additional input modality of spatially aligned and temporal
ly synchronized RGB and thermal cameras which enhances the accuracy of multi-cam
era tracking. MTMMC is a super-set of existing datasets benefiting independent f
ields such as person detection re-identification and multiple object tracking. W
e provide baselines and new learning setups on this dataset and set the referenc
e scores for future studies. The datasets models and test server will be made pu
blicly available.
****
Enhanced Motion-Text Alignment for Image-to-Video Transfer Learning
Wei Zhang, Chaoqun Wan, Tongliang Liu, Xinmei Tian, Xu Shen, Jieping Ye; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 18504-18515
Extending large image-text pre-trained models (e.g. CLIP) for video understandin
g has made significant advancements. To enable the capability of CLIP to perceiv
e dynamic information in videos existing works are dedicated to equipping the vi
sual encoder with various temporal modules. However these methods exhibit "asymm
etry" between the visual and textual sides with neither temporal descriptions in

input texts nor temporal modules in text encoder. This limitation hinders the potential of language supervision emphasized in CLIP and restricts the learning of temporal features as the text encoder has demonstrated limited proficiency in motion understanding. To address this issue we propose leveraging "MoTion-Enhanced Descriptions" (MoTED) to facilitate the extraction of distinctive temporal features in videos. Specifically we first generate discriminative motion-related descriptions via querying GPT-4 to compare easy-confusing action categories. Then we incorporate both the visual and textual encoders with additional perception modules to process the video frames and generated descriptions respectively. Finally we adopt a contrastive loss to align the visual and textual motion features. Extensive experiments on five benchmarks show that MoTED surpasses state-of-the-art methods with convincing gaps laying a solid foundation for empowering CLIP with strong temporal modeling.
****
DAP: A Dynamic Adversarial Patch for Evading Person Detectors
Amira Guesmi, Ruitian Ding, Muhammad Abdullah Hanif, Ihsen Alouani, Muhammad Shafique; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24595-24604
Patch-based adversarial attacks were proven to compromise the robustness and reliability of computer vision systems. However their conspicuous and easily detectable nature challenge their practicality in real-world setting. To address this recent work has proposed using Generative Adversarial Networks (GANs) to generate naturalistic patches that may not attract human attention. However such approaches suffer from a limited latent space making it challenging to produce a patch that is efficient stealthy and robust to multiple real-world transformations. This paper introduces a novel approach that produces a Dynamic Adversarial Patch (DAP) designed to overcome these limitations. DAP maintains a naturalistic appearance while optimizing attack efficiency and robustness to real-world transformations. The approach involves redefining the optimization problem and introducing a novel objective function that incorporates a similarity metric to guide the patch's creation. Unlike GAN-based techniques the DAP directly modifies pixel values within the patch providing increased flexibility and adaptability to multiple transformations. Furthermore most clothing-based physical attacks assume static objects and ignore the possible transformations caused by non-rigid deformation due to changes in a person's pose. To address this limitation a `Creases Transformation' (CT) block is introduced enhancing the patch's resilience to a variety of real-world distortions. Experimental results demonstrate that the proposed approach outperforms state-of-the-art attacks achieving a success rate of up to 82.28% in the digital world when targeting the YOLOv7 detector and 65% in the physical world when targeting YOLOv3tiny detector deployed in edge-based smart cameras.
****
Learned Lossless Image Compression based on Bit Plane Slicing
Zhe Zhang, Huairui Wang, Zhenzhong Chen, Shan Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27579-27588
Autoregressive Initial Bits (ArIB) a framework that combines subimage autoregression and latent variable models has shown its advantages in lossless image compression. However in current methods the image splitting makes the information of latent variables being uniformly distributed in each subimage and causes inadequate use of latent variables in addition to posterior collapse. To tackle these issues we introduce Bit Plane Slicing (BPS) splitting images in the bit plane dimension with the considerations on different importance for latent variables. Thus BPS provides a more effective representation by arranging subimages with decreasing importance for latent variables. To solve the problem of the increased number of dimensions caused by BPS we further propose a dimension-tailored autoregressive model that tailors autoregression methods for each dimension based on their characteristics efficiently capturing the dependencies in plane space and color dimensions. As shown in the extensive experimental results our method demonstrates the superior compression performance with comparable inference speed when

compared to the state-of-the-art normalizing-flow-based methods. The code is at https://github.com/ZZ022/ArIB-BPS.
****
UV-IDM: Identity-Conditioned Latent Diffusion Model for Face UV-Texture Generation
Hong Li, Yutang Feng, Song Xue, Xuhui Liu, Bohan Zeng, Shanglin Li, Boyu Liu, Jianzhuang Liu, Shumin Han, Baochang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10585-10595
3D face reconstruction aims at generating high-fidelity 3D face shapes and textures from single-view or multi-view images. However current prevailing facial texture generation methods generally suffer from low-quality texture identity information loss and inadequate handling of occlusions. To solve these problems we introduce an Identity-Conditioned Latent Diffusion Model for face UV-texture generation (UV-IDM) to generate photo-realistic textures based on the Basel Face Model (BFM). UV-IDM leverages the powerful texture generation capacity of a latent diffusion model (LDM) to obtain detailed facial textures. To preserve the identity during the reconstruction procedure we design an identity-conditioned module that can utilize any in-the-wild image as a robust condition for the LDM to guide texture generation. UV-IDM can be easily adapted to different BFM-based methods as a high-fidelity texture generator. Furthermore in light of the limited accessibility of most existing UV-texture datasets we build a large-scale and publicly available UV-texture dataset based on BFM termed BFM-UV. Extensive experiments show that our UV-IDM can generate high-fidelity textures in 3D face reconstruction within seconds while maintaining image consistency bringing new state-of-the-art performance in facial texture generation.
****
Mosaic-SDF for 3D Generative Models
Lior Yariv, Omri Puny, Oran Gafni, Yaron Lipman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4630-4639
Current diffusion or flow-based generative models for 3D shapes divide to two: distilling pre-trained 2D image diffusion models and training directly on 3D shapes. When training a diffusion or flow models on 3D shapes a crucial design choice is the shape representation. An effective shape representation needs to adhere three design principles: it should allow an efficient conversion of large 3D datasets to the representation form; it should provide a good tradeoff of approximation power versus number of parameters; and it should have a simple tensorial form that is compatible with existing powerful neural architectures. While standard 3D shape representations such as volumetric grids and point clouds do not adhere to all these principles simultaneously we advocate in this paper a new representation that does. We introduce Mosaic-SDF (M-SDF): a simple 3D shape representation that approximates the Signed Distance Function (SDF) of a given shape by using a set of local grids spread near the shape's boundary. The M-SDF representation is fast to compute for each shape individually making it readily parallelizable; it is parameter efficient as it only covers the space around the shape's boundary; and it has a simple matrix form compatible with Transformer-based architectures. We demonstrate the efficacy of the M-SDF representation by using it to train a 3D generative flow model including class-conditioned generation with the ShapeNetCore-V2 (3D Warehouse) dataset and text-to-3D generation using a dataset of about 600k caption-shape pairs.
****
Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D
Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, Niloy J. Mitra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7695-7704
Diffusion handles is a novel approach to enable 3D object edits on diffusion images requiring only existing pre-trained diffusion models depth estimation without any fine-tuning or 3D object retrieval. The edited results remain plausible photo-real and preserve object identity. Diffusion handles address a critically missing facet of generative image-based creative design. Our key insight is to lif

t diffusion activations for a selected object to 3D using a proxy depth 3D-trans form the depth and associated activations and project them back to image space. The diffusion process guided by the manipulated activations produces plausible e dited images showing complex 3D occlusion and lighting effects. We evaluate diff usion handles: quantitatively on a large synthetic data benchmark; and qualitati vely by a user study showing our output to be more plausible and better than pri or art at both 3D editing and identity control.
****
A Pedestrian is Worth One Prompt: Towards Language Guidance Person Re-Identifica tion
Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, Weiping Wang; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 17343-17353
Extensive advancements have been made in person ReID through the mining of seman tic information. Nevertheless existing methods that utilize semantic-parts from a single image modality do not explicitly achieve this goal. Whiteness the impre ssive capabilities in multimodal understanding of Vision Language Foundation Mod el CLIP a recent two-stage CLIP-based method employs automated prompt engineerin g to obtain specific textual labels for classifying pedestrians. However we note that the predefined soft prompts may be inadequate in expressing the entire vis ual context and struggle to generalize to unseen classes. This paper presents an end-to-end Prompt-driven Semantic Guidance (PromptSG) framework that harnesses the rich semantics inherent in CLIP. Specifically we guide the model to attend t o regions that are semantically faithful to the prompt. To provide personalized language descriptions for specific individuals we propose learning pseudo tokens that represent specific visual contexts. This design not only facilitates learn ing fine-grained attribute information but also can inherently leverage language prompts during inference. Without requiring additional labeling efforts our Pro mptSG achieves state-of-the-art by over 10% on MSMT17 and nearly 5% on the Marke t-1501 benchmark.
****
Friendly Sharpness-Aware Minimization
Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, Xiaolin Huang; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5631-5640
Sharpness-Aware Minimization (SAM) has been instrumental in improving deep neura l network training by minimizing both training loss and loss sharpness. Despite the practical success the mechanisms behind SAM's generalization enhancements re main elusive limiting its progress in deep learning optimization. In this work w e investigate SAM's core components for generalization improvement and introduce "Friendly-SAM" (F-SAM) to further enhance SAM's generalization. Our investigati on reveals the key role of batch-specific stochastic gradient noise within the a dversarial perturbation i.e. the current minibatch gradient which significantly influences SAM's generalization performance. By decomposing the adversarial pert urbation in SAM into full gradient and stochastic gradient noise components we d iscover that relying solely on the full gradient component degrades generalizati on while excluding it leads to improved performance. The possible reason lies in the full gradient component's increase in sharpness loss for the entire dataset creating inconsistencies with the subsequent sharpness minimization step solely on the current minibatch data. Inspired by these insights F-SAM aims to mitigat e the negative effects of the full gradient component. It removes the full gradi ent estimated by an exponentially moving average (EMA) of historical stochastic gradients and then leverages stochastic gradient noise for improved generalizati on. Moreover we provide theoretical validation for the EMA approximation and pro ve the convergence of F-SAM on non-convex problems. Extensive experiments demons trate the superior generalization performance and robustness of F-SAM over vanil la SAM. Code is available at https://github.com/nblt/F-SAM.
****
BIVDiff: A Training-Free Framework for General-Purpose Video Synthesis via Bridg ing Image and Video Diffusion Models

Fengyuan Shi, Jiaxi Gu, Hang Xu, Songcen Xu, Wei Zhang, Limin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7393-7402
Diffusion models have made tremendous progress in text-driven image and video generation. Now text-to-image foundation models are widely applied to various downstream image synthesis tasks such as controllable image generation and image editing while downstream video synthesis tasks are less explored for several reasons. First it requires huge memory and computation overhead to train a video generation foundation model. Even with video foundation models additional costly training is still required for downstream video synthesis tasks. Second although some works extend image diffusion models into videos in a training-free manner temporal consistency cannot be well preserved. Finally these adaption methods are specifically designed for one task and fail to generalize to different tasks. To mitigate these issues we propose a training-free general-purpose video synthesis framework coined as BIVDiff via bridging specific image diffusion models and general text-to-video foundation diffusion models. Specifically we first use a specific image diffusion model (e.g. ControlNet and Instruct Pix2Pix) for frame-wise video generation then perform Mixed Inversion on the generated video and finally input the inverted latents into the video diffusion models (e.g. VidRD and ZeroScope) for temporal smoothing. This decoupled framework enables flexible image model selection for different purposes with strong task generalization and high efficiency. To validate the effectiveness and general use of BIVDiff we perform a wide range of video synthesis tasks including controllable video generation video editing video inpainting and outpainting.
****
NC-TTT: A Noise Constrastive Approach for Test-Time Training
David Osowiechi, Gustavo A. Vargas Hakim, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, Christian Desrosiers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6078-6086
Despite their exceptional performance in vision tasks deep learning models often struggle when faced with domain shifts during testing. Test-Time Training (TTT) methods have recently gained popularity by their ability to enhance the robustness of models through the addition of an auxiliary objective that is jointly optimized with the main task. Being strictly unsupervised this auxiliary objective is used at test time to adapt the model without any access to labels. In this work we propose Noise-Contrastive Test-Time Training (NC-TTT) a novel unsupervised TTT technique based on the discrimination of noisy feature maps. By learning to classify noisy views of projected feature maps and then adapting the model accordingly on new domains classification performance can be recovered by an important margin. Experiments on several popular test-time adaptation baselines demonstrate the advantages of our method compared to recent approaches for this task. The code can be found at: https://github.com/GustavoVargasHakim/NCTTT.git
****
NetTrack: Tracking Highly Dynamic Objects with a Net
Guangze Zheng, Shijie Lin, Haobo Zuo, Changhong Fu, Jia Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19145-19155
The complex dynamicity of open-world objects presents non-negligible challenges for multi-object tracking (MOT) often manifested as severe deformations fast motion and occlusions. Most methods that solely depend on coarse-grained object cues such as boxes and the overall appearance of the object are susceptible to degradation due to distorted internal relationships of dynamic objects. To address this problem this work proposes NetTrack an efficient generic and affordable tracking framework to introduce fine-grained learning that is robust to dynamicity. Specifically NetTrack constructs a dynamicity-aware association with a fine-grained Net leveraging point-level visual cues. Correspondingly a fine-grained sampler and matching method have been incorporated. Furthermore NetTrack learns object-text correspondence for fine-grained localization. To evaluate MOT in extremely dynamic open-world scenarios a bird flock tracking (BFT) dataset is constructe

d which exhibits high dynamicity with diverse species and open-world scenarios. Comprehensive evaluation on BFT validates the effectiveness of fine-grained learning on object dynamicity and thorough transfer experiments on challenging open-world benchmarks i.e. TAO TAO-OW AnimalTrack and GMOT-40 validate the strong generalization ability of NetTrack even without finetuning.
****

Grounded Question-Answering in Long Egocentric Videos
Shangzhe Di, Weidi Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12934-12943
Existing approaches to video understanding mainly designed for short videos from a third-person perspective are limited in their applicability in certain fields such as robotics. In this paper we delve into open-ended question-answering (QA) in long egocentric videos which allows individuals or robots to inquire about their own past visual experiences. This task presents unique challenges including the complexity of temporally grounding queries within extensive video content the high resource demands for precise data annotation and the inherent difficulty of evaluating open-ended answers due to their ambiguous nature. Our proposed approach tackles these challenges by (i) integrating query grounding and answering within a unified model to reduce error propagation; (ii) employing large language models for efficient and scalable data synthesis; and (iii) introducing a close-ended QA task for evaluation to manage answer ambiguity. Extensive experiments demonstrate the effectiveness of our method which also achieves state-of-the-art performance on the QAEgo4D and Ego4D-NLQ benchmarks. Code data and models are open-sourced at https://github.com/Becomebright/GroundVQA.
****

HPNet: Dynamic Trajectory Forecasting with Historical Prediction Attention
Xiaolong Tang, Meina Kan, Shiguang Shan, Zhilong Ji, Jinfeng Bai, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15261-15270
Predicting the trajectories of road agents is essential for autonomous driving systems. The recent mainstream methods follow a static paradigm which predicts the future trajectory by using a fixed duration of historical frames. These methods make the predictions independently even at adjacent time steps which leads to potential instability and temporal inconsistency. As successive time steps have largely overlapping historical frames their forecasting should have intrinsic correlation such as overlapping predicted trajectories should be consistent or be different but share the same motion goal depending on the road situation. Motivated by this in this work we introduce HPNet a novel dynamic trajectory forecasting method. Aiming for stable and accurate trajectory forecasting our method leverages not only historical frames including maps and agent states but also historical predictions. Specifically we newly design a Historical Prediction Attention module to automatically encode the dynamic relationship between successive predictions. Besides it also extends the attention range beyond the currently visible window benefitting from the use of historical predictions. The proposed Historical Prediction Attention together with the Agent Attention and Mode Attention is further formulated as the Triple Factorized Attention module serving as the core design of HPNet. Experiments on the Argoverse and INTERACTION datasets show that HPNet achieves state-of-the-art performance and generates accurate and stable future trajectories. Our code are available at https://github.com/XiaolongTang23/HPNet.
****

Flexible Depth Completion for Sparse and Varying Point Densities
Jinhyung Park, Yu-Jhe Li, Kris Kitani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21540-21550
While recent depth completion methods have achieved remarkable results filling in relatively dense depth maps (e.g. projected 64-line LiDAR on KITTI or 500 sampled points on NYUv2) with RGB guidance their performance on very sparse input (e.g. 4-line LiDAR or 32 depth point measurements) is unverified. These sparser regimes present new challenges as a 4-line LiDAR increases the distance between pixels without depth and their nearest depth point sixfold from 5 pixels to 30 pix

els compared to 64 lines. Observing that existing methods struggle with sparse and variable distribution depth maps we propose an Affinity-Based Shift Correction (ASC) module that iteratively aligns depth predictions to input depth based on predicted affinities between image pixels and depth points. Our framework enables each depth point to adaptively influence and improve predictions across the image leading to largely improved results for fewer-line fewer-point and variable sparsity settings. Further we show improved performance in domain transfer from KITTI to nuScenes and from random sampling to irregular point distributions. Our correction module can easily be added to any depth completion or RGB-only depth estimation model notably allowing the latter to perform both completion and estimation with a single model.
****
Small Scale Data-Free Knowledge Distillation
He Liu, Yikai Wang, Huaping Liu, Fuchun Sun, Anbang Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6008-6016
Data-free knowledge distillation is able to utilize the knowledge learned by a large teacher network to augment the training of a smaller student network without accessing the original training data avoiding privacy security and proprietary risks in real applications. In this line of research existing methods typically follow an inversion-and-distillation paradigm in which a generative adversarial network on-the-fly trained with the guidance of the pre-trained teacher network is used to synthesize a large-scale sample set for knowledge distillation. In this paper we reexamine this common data-free knowledge distillation paradigm showing that there is considerable room to improve the overall training efficiency through a lens of "small-scale inverted data for knowledge distillation". In light of three empirical observations indicating the importance of how to balance class distributions in terms of synthetic sample diversity and difficulty during both data inversion and distillation processes we propose Small Scale Data-free Knowledge Distillation (SSD-KD). In formulation SSD-KD introduces a modulating function to balance synthetic samples and a priority sampling function to select proper samples facilitated by a dynamic replay buffer and a reinforcement learning strategy. As a result SSD-KD can perform distillation training conditioned on an extremely small scale of synthetic samples (e.g. 10x less than the original training data scale) making the overall training efficiency one or two orders of magnitude faster than many mainstream methods while retaining superior or competitive model performance as demonstrated on popular image classification and semantic segmentation benchmarks. The code is available at https://github.com/OSVAI/SSD-KD.
****
Shadows Don't Lie and Lines Can't Bend! Generative Models don't know Projective Geometry...for now
Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, D.A. Forsyth, Anand Bhattad; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28140-28149
Generative models can produce impressively realistic images. This paper demonstrates that generated images have geometric features different from those of real images. We build a set of collections of generated images prequalified to fool simple signal-based classifiers into believing they are real. We then show that prequalified generated images can be identified reliably by classifiers that only look at geometric properties. We use three such classifiers. All three classifiers are denied access to image pixels and look only at derived geometric features. The first classifier looks at the perspective field of the image the second looks at lines detected in the image and the third looks at relations between detected objects and shadows. Our procedure detects generated images more reliably than SOTA local signal based detectors for images from a number of distinct generators. Saliency maps suggest that the classifiers can identify geometric problems reliably. We conclude that current generators cannot reliably reproduce geometric properties of real images.
****

CFPL-FAS: Class Free Prompt Learning for Generalizable Face Anti-spoofing
Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, Zhen Lei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 222-232
Domain generalization (DG) based Face Anti-Spoofing (FAS) aims to improve the model's performance on unseen domains. Existing methods either rely on domain labels to align domain-invariant feature spaces or disentangle generalizable features from the whole sample which inevitably lead to the distortion of semantic feature structures and achieve limited generalization. In this work we make use of large-scale VLMs like CLIP and leverage the textual feature to dynamically adjust the classifier's weights for exploring generalizable visual features. Specifically we propose a novel Class Free Prompt Learning (CFPL) paradigm for DG FAS which utilizes two lightweight transformers namely Content Q-Former (CQF) and Style Q-Former (SQF) to learn the different semantic prompts conditioned on content and style features by using a set of learnable query vectors respectively. Thus the generalizable prompt can be learned by two improvements: (1) A Prompt-Text Matched (PTM) supervision is introduced to ensure CQF learns visual representation that is most informative of the content description. (2) A Diversified Style Prompt (DSP) technology is proposed to diversify the learning of style prompts by mixing feature statistics between instance-specific styles. Finally the learned text features modulate visual features to generalization through the designed Prompt Modulation (PM). Extensive experiments show that the CFPL is effective and outperforms the state-of-the-art methods on several cross-domain datasets.
****
SI-MIL: Taming Deep MIL for Self-Interpretability in Gigapixel Histopathology
Saarthak Kapse, Pushpak Pati, Srijan Das, Jingwei Zhang, Chao Chen, Maria Vakalopoulou, Joel Saltz, Dimitris Samaras, Rajarsi R. Gupta, Prateek Prasanna; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11226-11237
Introducing interpretability and reasoning into Multiple Instance Learning (MIL) methods for Whole Slide Image (WSI) analysis is challenging given the complexity of gigapixel slides. Traditionally MIL interpretability is limited to identifying salient regions deemed pertinent for downstream tasks offering little insight to the end-user (pathologist) regarding the rationale behind these selections. To address this we propose Self-Interpretable MIL (SI-MIL) a method intrinsically designed for interpretability from the very outset. SI-MIL employs a deep MIL framework to guide an interpretable branch grounded on handcrafted pathological features facilitating linear predictions. Beyond identifying salient regions SI-MIL uniquely provides feature-level interpretations rooted in pathological insights for WSIs. Notably SI-MIL with its linear prediction constraints challenges the prevalent myth of an inevitable trade-off between model interpretability and performance demonstrating competitive results compared to state-of-the-art methods on WSI-level prediction tasks across three cancer types. In addition we thoroughly benchmark the local- and global-interpretability of SI-MIL in terms of statistical analysis a domain expert study and desiderata of interpretability namely user-friendliness and faithfulness.
****
GEARS: Local Geometry-aware Hand-object Interaction Synthesis
Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20634-20643
Generating realistic hand motion sequences in interaction with objects has gained increasing attention with the growing interest in digital humans. Prior work has illustrated the effectiveness of employing occupancy-based or distance-based virtual sensors to extract hand-object interaction features. Nonetheless these methods show limited generalizability across object categories shapes and sizes. We hypothesize that this is due to two reasons: 1) the limited expressiveness of employed virtual sensors and 2) scarcity of available training data. To tackle this challenge we introduce a novel joint-centered sensor designed to reason about local object geometry near potential interaction regions. The sensor queries

for object surface points in the neighbourhood of each hand joint. As an importa
nt step towards mitigating the learning complexity we transform the points from
global frame to hand template frame and use a shared module to process sensor fe
atures of each individual joint. This is followed by a spatio-temporal transform
er network aimed at capturing correlation among the joints in different dimensio
ns. Moreover we devise simple heuristic rules to augment the limited training se
quences with vast static hand grasping samples. This leads to a broader spectrum
 of grasping types observed during training in turn enhancing our model's genera
lization capability. We evaluate on two public datasets GRAB and InterCap where
our method shows superiority over baselines both quantitatively and perceptually
.
****
Open Vocabulary Semantic Scene Sketch Understanding
Ahmed Bourouis, Judith E. Fan, Yulia Gryaditskaya; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4176-4186
We study the underexplored but fundamental vision problem of machine understandi
ng of abstract freehand scene sketches. We introduce a sketch encoder that resul
ts in semantically-aware feature space which we evaluate by testing its performa
nce on a semantic sketch segmentation task. To train our model we rely only on t
he availability of bitmap sketches with their brief captions and do not require
any pixel-level annotations. To obtain generalization to a large set of sketches
 and categories we build on a vision transformer encoder pretrained with the CLI
P model. We freeze the text encoder and perform visual-prompt tuning of the visu
al encoder branch while introducing a set of critical modifications. Firstly we
augment the classical key-query (k-q) self-attention blocks with value-value (v-
v) self-attention blocks. Central to our model is a two-level hierarchical netwo
rk design that enables efficient semantic disentanglement: The first level ensur
es holistic scene sketch encoding and the second level focuses on individual cat
egories. We then in the second level of the hierarchy introduce a cross-attentio
n between textual and visual branches. Our method outperforms zero-shot CLIP pix
el accuracy of segmentation results by 37 points reaching an accuracy of 85.5% o
n the FS-COCO sketch dataset. Finally we conduct a user study that allows us to
identify further improvements needed over our method to reconcile machine and hu
man understanding of scene sketches.
****
IntrinsicAvatar: Physically Based Inverse Rendering of Dynamic Humans from Monoc
ular Videos via Explicit Ray Tracing
Shaofei Wang, Bozidar Antic, Andreas Geiger, Siyu Tang; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1877
-1888
We present IntrinsicAvatar a novel approach to recovering the intrinsic properti
es of clothed human avatars including geometry albedo material and environment l
ighting from only monocular videos. Recent advancements in human-based neural re
ndering have enabled high-quality geometry and appearance reconstruction of clot
hed humans from just monocular videos. However these methods bake intrinsic prop
erties such as albedo material and environment lighting into a single entangled
neural representation. On the other hand only a handful of works tackle the prob
lem of estimating geometry and disentangled appearance properties of clothed hum
ans from monocular videos. They usually achieve limited quality and disentanglem
ent due to approximations of secondary shading effects via learned MLPs. In this
 work we propose to model secondary shading effects explicitly via Monte-Carlo r
ay tracing. We model the rendering process of clothed humans as a volumetric sca
ttering process and combine ray tracing with body articulation. Our approach can
 recover high-quality geometry albedo material and lighting properties of clothe
d humans from a single monocular video without requiring supervised pre-training
 using ground truth materials. Furthermore since we explicitly model the volumet
ric scattering process and ray tracing our model naturally generalizes to novel
poses enabling animation of the reconstructed avatar in novel lighting condition
s.
****

Efficient Detection of Long Consistent Cycles and its Application to Distributed Synchronization
Shaohan Li, Yunpeng Shi, Gilad Lerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5260-5269
Group synchronization plays a crucial role in global pipelines for Structure from Motion (SfM). Its formulation is nonconvex and it is faced with highly corrupted measurements. Cycle consistency has been effective in addressing these challenges. However computationally efficient solutions are needed for cycles longer than three especially in practical scenarios where 3-cycles are unavailable. To overcome this computational bottleneck we propose an algorithm for group synchronization that leverages information from cycles of lengths ranging from three to six with a complexity of O(n^3) (or O(n^ 2.373) when using a faster matrix multiplication algorithm). We establish non-trivial theory for this and related methods that achieves competitive sample complexity assuming the uniform corruption model. To advocate the practical need for our method we consider distributed group synchronization which requires at least 4-cycles and we illustrate state-of-the-art performance by our method in this context.
****
LayoutFormer: Hierarchical Text Detection Towards Scene Text Understanding
Min Liang, Jia-Wei Ma, Xiaobin Zhu, Jingyan Qin, Xu-Cheng Yin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15665-15674
Existing scene text detectors generally focus on accurately detecting single-level (i.e. word-level line-level or paragraph-level) text entities without exploring the relationships among different levels of text entities. To comprehensively understand scene texts detecting multi-level texts while exploring their contextual information is critical. To this end we propose a unified framework (dubbed LayoutFormer) for hierarchical text detection which simultaneously conducts multi-level text detection and predicts the geometric layouts for promoting scene text understanding. In LayoutFormer WordDecoder LineDecoder and ParaDecoder are proposed to be responsible for word-level text prediction line-level text prediction and paragraph-level text prediction respectively. Meanwhile WordDecoder and ParaDecoder adaptively learn word-line and line-paragraph relationships respectively. In addition we propose a Prior Location Sampler to be used on multi-scale features to adaptively select a few representative foreground features for updating text queries. It can improve hierarchical detection performance while significantly reducing the computational cost. Comprehensive experiments verify that our method achieves state-of-the-art performance on single-level and hierarchical text detection.
****
Vlogger: Make Your Dream A Vlog
Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, Yali Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8806-8817
In this work we present Vlogger a generic AI system for generating a minute-level video blog (i.e. vlog) of user descriptions. Different from short videos with a few seconds vlog often contains a complex storyline with diversified scenes which is challenging for most existing video generation approaches. To break through this bottleneck our Vlogger smartly leverages Large Language Model (LLM) as Director and decomposes a long video generation task of vlog into four key stages where we invoke various foundation models to play the critical roles of vlog professionals including (1) Script (2) Actor (3) ShowMaker and (4) Voicer. With such a design of mimicking human beings our Vlogger can generate vlogs through explainable cooperation of top-down planning and bottom-up shooting. More over we introduce a novel video diffusion model ShowMaker which serves as a videographer in our Vlogger for generating the video snippet of each shooting scene. By incorporating Script and Actor attentively as textual and visual prompts it can effectively enhance spatial-temporal coherence in the snippet. Besides we design a concise mixed training paradigm for ShowMaker boosting its capacity for both T2V generation and prediction. Finally the extensive experiments show that our method

achieves state-of-the-art performance on zero-shot T2V generation and prediction tasks. More importantly Vlogger can generate over 5-minute vlogs from open-world descriptions without loss of video coherence on script and actor.
****
CodedEvents: Optimal Point-Spread-Function Engineering for 3D-Tracking with Event Cameras
Sachin Shah, Matthew A. Chan, Haoming Cai, Jingxi Chen, Sakshum Kulshrestha, Chahat Deep Singh, Yiannis Aloimonos, Christopher A. Metzler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25265-25275
Point-spread-function (PSF) engineering is a well-established computational imaging technique that uses phase masks and other optical elements to embed extra information (e.g. depth) into the images captured by conventional CMOS image sensors. To date however PSF-engineering has not been applied to neuromorphic event cameras; a powerful new image sensing technology that responds to changes in the log-intensity of light. This paper establishes theoretical limits (Cramer Rao bounds) on 3D point localization and tracking with PSF-engineered event cameras. Using these bounds we first demonstrate that existing Fisher phase masks are already near-optimal for localizing static flashing point sources (e.g. blinking fluorescent molecules). We then demonstrate that existing designs are sub-optimal for tracking moving point sources and proceed to use our theory to design optimal phase masks and binary amplitude masks for this task. To overcome the non-convexity of the design problem we leverage novel implicit neural representation based parameterizations of the phase and amplitude masks. We demonstrate the efficacy of our designs through extensive simulations. We also validate our method with a simple prototype.
****
GLOW: Global Layout Aware Attacks on Object Detection
Jun Bao, Buyu Liu, Kui Ren, Jun Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12057-12066
Adversarial attacks aim to perturb images such that a predictor outputs incorrect results. Due to the limited research in structured attacks imposing consistency checks on natural multi-object scenes is a practical defense against conventional adversarial attacks. More desired attacks should be able to fool defenses with such consistency checks. Therefore we present the first approach GLOW that copes with various attack requests by generating global layout-aware adversarial attacks in which both categorical and geometric layout constraints are explicitly established. Specifically we focus on object detection tasks and given a victim image GLOW first localizes victim objects according to target labels. And then it generates multiple attack plans together with their context-consistency scores. GLOW on the one hand is capable of handling various types of requests including single or multiple victim objects with or without specified victim objects. On the other hand it produces a consistency score for each attack plan reflecting the overall contextual consistency that both semantic category and global scene layout are considered. We conduct our experiments on MS COCO and Pascal. Extensive experimental results demonstrate that we can achieve about 30% average relative improvement compared to state-of-the-art methods in conventional single object attack request; Moreover such superiority is also valid across more generic attack requests under both white-box and zero-query black-box settings. Finally we conduct comprehensive human analysis which not only validates our claim further but also provides strong evidence that our evaluation metrics reflect human reviews well.
****
Learning Discriminative Dynamics with Label Corruption for Noisy Label Detection
Suyeon Kim, Dongha Lee, SeongKu Kang, Sukang Chae, Sanghwan Jang, Hwanjo Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22477-22487
Label noise commonly found in real-world datasets has a detrimental impact on a model's generalization. To effectively detect incorrectly labeled instances previous works have mostly relied on distinguishable training signals such as traini

ng loss as indicators to differentiate between clean and noisy labels. However they have limitations in that the training signals incompletely reveal the model's behavior and are not effectively generalized to various noise types resulting in limited detection accuracy. In this paper we propose DynaCor framework that distinguishes incorrectly labeled instances from correctly labeled ones based on the dynamics of the training signals. To cope with the absence of supervision for clean and noisy labels DynaCor first introduces a label corruption strategy that augments the original dataset with intentionally corrupted labels enabling indirect simulation of the model's behavior on noisy labels. Then DynaCor learns to identify clean and noisy instances by inducing two clearly distinguishable clusters from the latent representations of training dynamics. Our comprehensive experiments show that DynaCor outperforms the state-of-the-art competitors and shows strong robustness to various noise types and noise rates.
****
Neural 3D Strokes: Creating Stylized 3D Scenes with Vectorized 3D Strokes
Hao-Bin Duan, Miao Wang, Yan-Xun Li, Yong-Liang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5240-5249
We present Neural 3D Strokes a novel technique to generate stylized images of a 3D scene at arbitrary novel views from multi-view 2D images. Different from existing methods which apply stylization to trained neural radiance fields at the voxel level our approach draws inspiration from image-to-painting methods simulating the progressive painting process of human artwork with vector strokes. We develop a palette of stylized 3D strokes from basic primitives and splines and consider the 3D scene stylization task as a multi-view reconstruction process based on these 3D stroke primitives. Instead of directly searching for the parameters of these 3D strokes which would be too costly we introduce a differentiable renderer that allows optimizing stroke parameters using gradient descent and propose a training scheme to alleviate the vanishing gradient issue. The extensive evaluation demonstrates that our approach effectively synthesizes 3D scenes with significant geometric and aesthetic stylization while maintaining a consistent appearance across different views. Our method can be further integrated with style loss and image-text contrastive models to extend its applications including color transfer and text-driven 3D scene drawing. Results and code are available at http://buaavrcg.github.io/Neural3DStrokes.
****
SIRA: Scalable Inter-frame Relation and Association for Radar Perception
Ryoma Yataka, Pu Wang, Petros Boufounos, Ryuhei Takahashi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15024-15034
Conventional radar feature extraction faces limitations due to low spatial resolution noise multipath reflection the presence of ghost targets and motion blur. Such limitations can be exacerbated by nonlinear object motion particularly from an ego-centric viewpoint. It becomes evident that to address these challenges the key lies in exploiting temporal feature relation over an extended horizon and enforcing spatial motion consistence for effective association. To this end this paper proposes SIRA (Scalable Inter-frame Relation and Association) with two designs. First inspired by Swin Transformer we introduce extended temporal relation generalizing the existing temporal relation layer from two consecutive frames to multiple inter-frames with temporally regrouped window attention for scalability. Second we propose motion consistency track with the concept of a pseudo-tracklet generated from observational data for better trajectory prediction and subsequent object association. Our approach achieves 58.11 mAP@0.5 for oriented object detection and 47.79 MOTA for multiple object tracking on the Radiate dataset surpassing previous state-of-the-art by a margin of +4.11 mAP@0.5 and +9.94 MOTA respectively.
****
VOODOO 3D: Volumetric Portrait Disentanglement For One-Shot 3D Head Reenactment
Phong Tran, Egor Zakharov, Long-Nhat Ho, Anh Tuan Tran, Liwen Hu, Hao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP

R), 2024, pp. 10336-10348
We present a 3D-aware one-shot head reenactment method based on a fully volumetric neural disentanglement framework for source appearance and driver expressions. Our method is real-time and produces high-fidelity and view-consistent output suitable for 3D teleconferencing systems based on holographic displays. Existing cutting-edge 3D-aware reenactment methods often use neural radiance fields or 3D meshes to produce view-consistent appearance encoding but at the same time they rely on linear face models such as 3DMM to achieve its disentanglement with facial expressions. As a result their reenactment results often exhibit identity leakage from the driver or have unnatural expressions. To address these problems we propose a neural self-supervised disentanglement approach that lifts both the source image and driver video frame into a shared 3D volumetric representation based on tri-planes. This representation can then be freely manipulated with expression tri-planes extracted from the driving images and rendered from an arbitrary view using neural radiance fields. We achieve this disentanglement via self-supervised learning on a large in-the-wild video dataset. We further introduce a highly effective fine-tuning approach to improve the generalizability of the 3D lifting using the same real-world data. We demonstrate state-of-the-art performance on a wide range of datasets and also showcase high-quality 3D-aware head reenactment on highly challenging and diverse subjects including non-frontal head poses and complex expressions for both source and driver.
****
Visual Fact Checker: Enabling High-Fidelity Detailed Caption Generation
Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, Yin Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14033-14042
Existing automatic captioning methods for visual content face challenges such as lack of detail content hallucination and poor instruction following. In this work we propose VisualFactChecker (VFC) a flexible training-free pipeline that generates high-fidelity and detailed captions for both 2D images and 3D objects. VFC consists of three steps: 1) proposal where image-to-text captioning models propose multiple initial captions; 2) verification where a large language model (LLM) utilizes tools such as object detection and VQA models to fact-check proposed captions; 3) captioning where an LLM generates the final caption by summarizing caption proposals and the fact check verification results. In this step VFC can flexibly generate captions in various styles following complex instructions. We conduct comprehensive captioning evaluations using four metrics: 1) CLIP-Score for image-text similarity; 2) CLIP-Image-Score for measuring the image-image similarity between the original and the reconstructed image generated by a text-to-image model using the caption. 3) human study on Amazon Mechanical Turk; 4) GPT-4V for fine-grained evaluation. Evaluation results show that VFC outperforms state-of-the-art open-sourced captioning methods for 2D images on the COCO dataset and 3D assets on the Objaverse dataset. Our study demonstrates that by combining open-source models into a pipeline we can attain captioning capability comparable to proprietary models such as GPT-4V despite being over 10x smaller in model size.
****
Communication-Efficient Collaborative Perception via Information Filling with Codebook
Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, Siheng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15481-15490
Collaborative perception empowers each agent to improve its perceptual ability through the exchange of perceptual messages with other agents. It inherently results in a fundamental trade-off between perception ability and communication cost. To address this bottleneck issue our core idea is to optimize the collaborative messages from two key aspects: representation and selection. The proposed codebook-based message representation enables the transmission of integer codes rather than high-dimensional feature maps. The proposed information-filling-driven message selection optimizes local messages to collectively fill each agent's info

rmation demand preventing information overflow among multiple agents. By integrating these two designs we propose CodeFilling a novel communication-efficient collaborative perception system which significantly advances the perception-communication trade-off and is inclusive to both homogeneous and heterogeneous collaboration settings. We evaluate CodeFilling in both a real-world dataset DAIR-V2X and a new simulation dataset OPV2VH+. Results show that CodeFilling outperforms previous SOTA Where2comm on DAIR-V2X/OPV2VH+ with 1333/1206x lower communication volume. Our code is available at https://github.com/PhyllisH/CodeFilling.
****
DiPrompT: Disentangled Prompt Tuning for Multiple Latent Domain Generalization in Federated Learning
Sikai Bai, Jie Zhang, Song Guo, Shuaicheng Li, Jingcai Guo, Jun Hou, Tao Han, Xiaocheng Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27284-27293
Federated learning (FL) has emerged as a powerful paradigm for learning from decentralized data and federated domain generalization further considers the test dataset (target domain) is absent from the decentralized training data (source domains). However most existing FL methods assume that domain labels are provided during training and their evaluation imposes explicit constraints on the number of domains which must strictly match the number of clients. Because of the underutilization of numerous edge devices and additional cross-client domain annotations in the real world such restrictions may be impractical and involve potential privacy leaks. In this paper we propose an efficient and novel approach called Disentangled Prompt Tuning (DiPrompT) a method that tackles the above restrictions by learning adaptive prompts for domain generalization in a distributed manner. Specifically we first design two types of prompts i.e. global prompt to capture general knowledge across all clients and domain prompts to capture domain-specific knowledge. They eliminate the restriction on the one-to-one mapping between source domains and local clients. Furthermore a dynamic query metric is introduced to automatically search the suitable domain label for each sample which includes two-substep text-image alignments based on prompt tuning without labor-intensive annotation. Extensive experiments on multiple datasets demonstrate that our DiPrompT achieves superior domain generalization performance over state-of-the-art FL methods when domain labels are not provided and even outperforms many centralized learning methods using domain labels.
****
MVD-Fusion: Single-view 3D via Depth-consistent Multi-view Generation
Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, Shubham Tulsiani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9698-9707
We present MVD-Fusion: a method for single-view 3D inference via generative modeling of multi-view-consistent RGB-D images. While recent methods pursuing 3D inference advocate learning novel-view generative models these generations are not 3D-consistent and require a distillation process to generate a 3D output. We instead cast the task of 3D inference as directly generating mutually-consistent multiple views and build on the insight that additionally inferring depth can provide a mechanism for enforcing this consistency. Specifically we train a denoising diffusion model to generate multi-view RGB-D images given a single RGB input image and leverage the (intermediate noisy) depth estimates to obtain reprojection-based conditioning to maintain multi-view consistency. We train our model using large-scale synthetic dataset Obajverse as well as the real-world CO3D dataset comprising of generic camera viewpoints. We demonstrate that our approach can yield more accurate synthesis compared to recent state-of-the-art including distillation-based 3D inference and prior multi-view generation methods. We also evaluate the geometry induced by our multi-view depth prediction and find that it yields a more accurate representation than other direct 3D inference approaches.
****
Effective Video Mirror Detection with Inconsistent Motion Cues
Alex Warren, Ke Xu, Jiaying Lin, Gary K.L. Tam, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,

pp. 17244-17252
Image-based mirror detection has recently undergone rapid research due to its si gnificance in applications such as robotic navigation semantic segmentation and scene reconstruction. Recently VMD-Net was proposed as the first video mirror de tection technique by modeling dual correspondences between the inside and outsid e of the mirror both spatially and temporally. However this approach is not reli able as correspondences can occur completely inside or outside of the mirrors. I n addition the proposed dataset VMD-D contains many small mirrors limiting its a pplicability to real-world scenarios. To address these problems we developed a m ore challenging dataset that includes mirrors of various shapes and sizes at dif ferent locations of the frames providing a better reflection of real-world scena rios. Next we observed that the motions between the inside and outside of the mi rror are often inconsistent. For instance when moving in front of a mirror the m otion inside the mirror is often much smaller than the motion outside due to inc reased depth perception. With these observations we propose modeling inconsisten t motion cues to detect mirrors and a new network with two novel modules. The Mo tion Attention Module (MAM) explicitly models inconsistent motions around mirror s via optical flow and the Motion-Guided Edge Detection Module (MEDM) uses motio ns to guide mirror edge feature learning. Experimental results on our proposed d ataset show that our method outperforms state-of-the-arts. The code and dataset are available at https://github.com/AlexAnthonyWarren/MG-VMD.
****
Multi-Object Tracking in the Dark
Xinzhe Wang, Kang Ma, Qiankun Liu, Yunhao Zou, Ying Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 382-392
Low-light scenes are prevalent in real-world applications (e.g. autonomous drivi ng and surveillance at night). Recently multi-object tracking in various practic al use cases have received much attention but multi-object tracking in dark scen es is rarely considered. In this paper we focus on multi-object tracking in dark scenes. To address the lack of datasets we first build a Low-light Multi-Object Tracking (LMOT) dataset. LMOT provides well-aligned low-light video pairs captu red by our dual-camera system and high-quality multi-object tracking annotations for all videos. Then we propose a low-light multi-object tracking method termed as LTrack. We introduce the adaptive low-pass downsample module to enhance low-frequency components of images outside the sensor noises. The degradation suppre ssion learning strategy enables the model to learn invariant information under n oise disturbance and image quality degradation. These components improve the rob ustness of multi-object tracking in dark scenes. We conducted a comprehensive an alysis of our LMOT dataset and proposed LTrack. Experimental results demonstrate the superiority of the proposed method and its competitiveness in real night lo w-light scenes. Dataset and Code: https://github.com/ying-fu/LMOT
****
UniHuman: A Unified Model For Editing Human Images in the Wild
Nannan Li, Qing Liu, Krishna Kumar Singh, Yilin Wang, Jianming Zhang, Bryan A. Plummer, Zhe Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2039-2048
Human image editing includes tasks like changing a person's pose their clothing or editing the image according to a text prompt. However prior work often tackle s these tasks separately overlooking the benefit of mutual reinforcement from le arning them jointly. In this paper we propose UniHuman a unified model that addr esses multiple facets of human image editing in real-world settings. To enhance the model's generation quality and generalization capacity we leverage guidance from human visual encoders and introduce a lightweight pose-warping module that can exploit different pose representations accommodating unseen textures and pat terns. Furthermore to bridge the disparity between existing human editing benchm arks with real-world data we curated 400K high-quality human image-text pairs fo r training and collected 2K human images for out-of-domain testing both encompas sing diverse clothing styles backgrounds and age groups. Experiments on both in-domain and out-of-domain test sets demonstrate that UniHuman outperforms task-sp

ecific models by a significant margin. In user studies UniHuman is preferred by the users in an average of 77% of cases. Our project is available at https://github.com/NannanLi999/UniHuman.
****
DiffAgent: Fast and Accurate Text-to-Image API Selection with Large Language Model
Lirui Zhao, Yue Yang, Kaipeng Zhang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6390-6399
Text-to-image (T2I) generative models have attracted significant attention and found extensive applications within and beyond academic research. For example the Civitai community a platform for T2I innovation currently hosts an impressive array of 74492 distinct models. However this diversity presents a formidable challenge in selecting the most appropriate model and parameters a process that typically requires numerous trials. Drawing inspiration from the tool usage research of large language models (LLMs) we introduce DiffAgent an LLM agent designed to screen the accurate selection in seconds via API calls. DiffAgent leverages a novel two-stage training framework SFTA enabling it to accurately align T2I API responses with user input in accordance with human preferences. To train and evaluate DiffAgent's capabilities we present DABench a comprehensive dataset encompassing an extensive range of T2I APIs from the community. Our evaluations reveal that DiffAgent not only excels in identifying the appropriate T2I API but also underscores the effectiveness of the SFTA training framework. Codes are available at https://github.com/OpenGVLab/DiffAgent.
****
In Search of a Data Transformation That Accelerates Neural Field Training
Junwon Seo, Sangyoon Lee, Kwang In Kim, Jaeho Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4830-4839
Neural field is an emerging paradigm in data representation that trains a neural network to approximate the given signal. A key obstacle that prevents its widespread adoption is the encoding speed---generating neural fields requires an overfitting of a neural network which can take a significant number of SGD steps to reach the desired fidelity level. In this paper we delve into the impacts of data transformations on the speed of neural field training specifically focusing on how permuting pixel locations affect the convergence speed of SGD. Counterintuitively we find that randomly permuting the pixel locations can considerably accelerate the training. To explain this phenomenon we examine the neural field training through the lens of PSNR curves loss landscapes and error patterns. Our analyses suggest that the random pixel permutations remove the easy-to-fit patterns which facilitate easy optimization in the early stage but hinder capturing fine details of the signal.
****
Zero-Painter: Training-Free Layout Control for Text-to-Image Synthesis
Marianna Ohanyan, Hayk Manukyan, Zhangyang Wang, Shant Navasardyan, Humphrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8764-8774
We present Zero-Painter a novel training-free framework for layout-conditional text-to-image synthesis that facilitates the creation of detailed and controlled imagery from textual prompts. Our method utilizes object masks and individual descriptions coupled with a global text prompt to generate images with high fidelity. Zero-Painter employs a two-stage process involving our novel Prompt-Adjusted Cross-Attention (PACA) and Region-Grouped Cross-Attention (ReGCA) blocks ensuring precise alignment of generated objects with textual prompts and mask shapes. Our extensive experiments demonstrate that Zero-Painter surpasses current state-of-the-art methods in preserving textual details and adhering to mask shapes. We will make the codes and the models publicly available.
****
DiffLoc: Diffusion Model for Outdoor LiDAR Localization
Wen Li, Yuyang Yang, Shangshu Yu, Guosheng Hu, Chenglu Wen, Ming Cheng, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn

ition (CVPR), 2024, pp. 15045-15054
Absolute pose regression (APR) estimates global pose in an end-to-end manner achieving impressive results in learn-based LiDAR localization. However compared to the top-performing methods reliant on 3D-3D correspondence matching APR's accuracy still has room for improvement. We recognize APR's lack of robust features learning and iterative denoising process leads to suboptimal results. In this paper we propose DiffLoc a novel framework that formulates LiDAR localization as a conditional generation of poses. First we propose to utilize the foundation model and static-object-aware pool to learn robust features. Second we incorporate the iterative denoising process into APR via a diffusion model conditioned on the learned geometrically robust features. In addition due to the unique nature of diffusion models we propose to adapt our models to two additional applications: (1) using multiple inferences to evaluate pose uncertainty and (2) seamlessly introducing geometric constraints on denoising steps to improve prediction accuracy. Extensive experiments conducted on the Oxford Radar RobotCar and NCLT datasets demonstrate that DiffLoc outperforms better than the stateof-the-art methods. Especially on the NCLT dataset we achieve 35% and 34.7% improvement on position and orientation accuracy respectively. Our code is released at https://github.com/liw95/DiffLoc.
****
Towards 3D Vision with Low-Cost Single-Photon Cameras
Fangzhou Mu, Carter Sifferman, Sacha Jungerman, Yiquan Li, Mark Han, Michael Gleicher, Mohit Gupta, Yin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5302-5311
We present a method for reconstructing 3D shape of arbitrary Lambertian objects based on measurements by miniature energy-efficient low-cost single-photon cameras. These cameras operating as time resolved image sensors illuminate the scene with a very fast pulse of diffuse light and record the shape of that pulse as it returns back from the scene at a high temporal resolution. We propose to model this image formation process account for its non-idealities and adapt neural rendering to reconstruct 3D geometry from a set of spatially distributed sensors with known poses. We show that our approach can successfully recover complex 3D shapes from simulated data. We further demonstrate 3D object reconstruction from real-world captures utilizing measurements from a commodity proximity sensor. Our work draws a connection between image-based modeling and active range scanning and offers a step towards 3D vision with single-photon cameras.
****
WonderJourney: Going from Anywhere to Everywhere
Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, Charles Herrmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6658-6667
We introduce WonderJourney a modular framework for perpetual 3D scene generation. Unlike prior work on view generation that focuses on a single type of scenes we start at any user-provided location (by a text description or an image) and generate a journey through a long sequence of diverse yet coherently connected 3D scenes. We leverage an LLM to generate textual descriptions of the scenes in this journey a text-driven point cloud generation pipeline to make a compelling and coherent sequence of 3D scenes and a large VLM to verify the generated scenes. We show compelling diverse visual results across various scene types and styles forming imaginary "wonderjourneys". Project website: https://kovenyu.com/WonderJourney.
****
On Scaling Up a Multilingual Vision and Language Model
Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter

Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, Radu Soricut; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14432-14444

We explore the boundaries of scaling up a multilingual vision and language model both in terms of size of the components and the breadth of its training task mixture. Our model achieves new levels of performance on a wide-range of varied and complex tasks including multiple image-based captioning and question-answering tasks image-based document understanding and few-shot (in-context) learning as well as object detection video question answering and video captioning. Our model advances the state-of-the-art on most vision-and-language benchmarks considered (20+ of them). Finally we observe emerging capabilities such as complex counting and multilingual object detection tasks that are not explicitly in the training mix.

****

Day-Night Cross-domain Vehicle Re-identification

Hongchao Li, Jingong Chen, Aihua Zheng, Yong Wu, Yonglong Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12626-12635

Previous advances in vehicle re-identification (ReID) are mostly reported under favorable lighting conditions while cross-day-and-night performance is neglected which greatly hinders the development of related traffic intelligence applications. This work instead develops a novel Day-Night Dual-domain Modulation (DNDM) vehicle re-identification framework for day-night cross-domain traffic scenarios. Specifically a unique night-domain glare suppression module is provided to attenuate the headlight glare from raw nighttime vehicle images. To enhance vehicle features under low-light environments we propose a dual-domain structure enhancement module in the feature extractor which enhances geometric structures between appearance features. To alleviate day-night domain discrepancies we develop a cross-domain class awareness module that facilitates the interaction between appearance and structure features in both domains. In this work we address the Day-Night cross-domain ReID (DN-ReID) problem and provide a new cross-domain dataset named DN-Wild including day and night images of 2286 identities giving in total 85945 daytime images and 54952 nighttime images. Furthermore we also take into account the matter of balance between day and night samples and provide a dataset called DN-348. Exhaustive experiments demonstrate the robustness of the proposed framework in the DN-ReID problem. The code and benchmark are released at https://github.com/chenjingong/DN-ReID.

****

4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, David B. Lindell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7996-8006

Recent breakthroughs in text-to-4D generation rely on pre-trained text-to-image and text-to-video models to generate dynamic 3D scenes. However current text-to-4D methods face a three-way tradeoff between the quality of scene appearance 3D structure and motion. For example text-to-image models and their 3D-aware variants are trained on internet-scale image datasets and can be used to produce scenes with realistic appearance and 3D structure---but no motion. Text-to-video models are trained on relatively smaller video datasets and can produce scenes with motion but poorer appearance and 3D structure. While these models have complementary strengths they also have opposing weaknesses making it difficult to combine them in a way that alleviates this three-way tradeoff. Here we introduce hybrid score distillation sampling an alternating optimization procedure that blends supervision signals from multiple pre-trained diffusion models and incorporates benefits of each for high-fidelity text-to-4D generation. Using hybrid SDS we demonstrate synthesis of 4D scenes with compelling appearance 3D structure and motion.

****

Adversarial Distillation Based on Slack Matching and Attribution Region Alignmen
t
Shenglin Yin, Zhen Xiao, Mingxuan Song, Jieyi Long; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24605-24
614
Adversarial distillation (AD) is a highly effective method for enhancing the rob
ustness of small models. Contrary to expectations a high-performing teacher mode
l does not always result in a more robust student model. This is due to two main
 reasons. First when there are significant differences in predictions between th
e teacher model and the student model exact matching of predicted values using K
L divergence interferes with training leading to poor performance of existing me
thods. Second matching solely based on the output prevents the student model fro
m fully understanding the behavior of the teacher model. To address these challe
nges this paper proposes a novel AD method named SmaraAD. During the training pr
ocess we facilitate the student model in better understanding the teacher model'
s behavior by aligning the attribution region that the student model focuses on
with that of the teacher model. Concurrently we relax the condition of exact mat
ching in KL divergence and replace it with a more flexible matching criterion th
ereby enhancing the model's robustness. Extensive experiments substantiate the e
ffectiveness of our method in improving the robustness of small models outperfor
ming previous SOTA methods.
****
Boosting Spike Camera Image Reconstruction from a Perspective of Dealing with Sp
ike Fluctuations
Rui Zhao, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Zhaofei Yu, Tiejun
Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2024, pp. 24955-24965
As a bio-inspired vision sensor with ultra-high speed spike cameras exhibit grea
t potential in recording dynamic scenes with high-speed motion or drastic light
changes. Different from traditional cameras each pixel in spike cameras records
the arrival of photons continuously by firing binary spikes at an ultra-fine tem
poral granularity. In this process multiple factors impact the imaging including
 the photons' Poisson arrival thermal noises from circuits and quantization effe
cts in spike readout. These factors introduce fluctuations to spikes making the
recorded spike intervals unstable and unable to reflect accurate light intensiti
es. In this paper we present an approach to deal with spike fluctuations and boo
st spike camera image reconstruction. We first analyze the quantization effects
and reveal the unbiased estimation attribute of the reciprocal of differential o
f spike firing time (DSFT). Based on this we propose a spike representation modu
le to use DSFT with multiple orders for fluctuation suppression where DSFT with
higher orders indicates spike integration duration between multiple spikes. We a
lso propose a module for inter-moment feature alignment at multiple granularitie
s. The coarser alignment is based on patch-level cross-attention with a local se
arch strategy and the finer alignment is based on deformable convolution at the
pixel level. Experimental results demonstrate the effectiveness of our method on
 both synthetic and real-captured data. The source code and dataset are availabl
e at https://github.com/ruizhao26/BSF.
****
Text-guided Explorable Image Super-resolution
Kanchana Vaishnavi Gandikota, Paramanand Chandramouli; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25900
-25911
In this paper we introduce the problem of zero-shot text-guided exploration of t
he solutions to open-domain image super-resolution. Our goal is to allow users t
o explore diverse semantically accurate reconstructions that preserve data consi
stency with the low-resolution inputs for different large downsampling factors w
ithout explicitly training for these specific degradations. We propose two appro
aches for zero-shot text-guided super-resolution - i) modifying the generative p
rocess of text-to-image (T2I) diffusion models to promote consistency with low-r
esolution inputs and ii) incorporating language guidance into zero-shot diffusio

n-based restoration methods. We show that the proposed approaches result in diverse solutions that match the semantic meaning provided by the text prompt while preserving data consistency with the degraded inputs. We evaluate the proposed baselines for the task of extreme super-resolution and demonstrate advantages in terms of restoration quality diversity and explorability of solutions.
****
FreeControl: Training-Free Spatial Control of Any Text-to-Image Diffusion Model with Any Condition
Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7465-7475
Recent approaches such as ControlNet offer users fine-grained spatial control over text-to-image (T2I) diffusion models. However auxiliary modules have to be trained for each spatial condition type model architecture and checkpoint putting them at odds with the diverse intents and preferences a human designer would like to convey to the AI models during the content creation process. In this work we present FreeControl a training-free approach for controllable T2I generation that supports multiple conditions architectures and checkpoints simultaneously. FreeControl enforces structure guidance to facilitate the global alignment with a guidance image and appearance guidance to collect visual details from images generated without control. Extensive qualitative and quantitative experiments demonstrate the superior performance of FreeControl across a variety of pre-trained T2I models. In particular FreeControl enables convenient training-free control over many different architectures and checkpoints allows the challenging input conditions on which most of the existing training-free methods fail and achieves competitive synthesis quality compared to training-based approaches. Project page :https://genforce.github.io/freecontrol/.
****
VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models
Hyeonho Jeong, Geon Yeong Park, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9212-9221
Text-to-video diffusion models have advanced video generation significantly. However customizing these models to generate videos with tailored motions presents a substantial challenge. In specific they encounter hurdles in (1) accurately reproducing motion from a target video and (2) creating diverse visual variations. For example straightforward extensions of static image customization methods to video often lead to intricate entanglements of appearance and motion data. To tackle this here we present the Video Motion Customization (VMC) framework a novel one-shot tuning approach crafted to adapt temporal attention layers within video diffusion models. Our approach introduces a novel motion distillation objective using residual vectors between consecutive noisy latent frames as a motion reference. The diffusion process then preserve low-frequency motion trajectories while mitigating high-frequency motion-unrelated noise in image space. We validate our method against state-of-the-art video generative models across diverse real-world motions and contexts. Our codes and data can be found at: https://video-motion-customization.github.io/
****
Holodeck: Language Guided Generation of 3D Embodied AI Environments
Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, Christopher Clark; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16227-16237
3D simulated environments play a critical role in Embodied AI but their creation requires expertise and extensive manual effort restricting their diversity and scope. To mitigate this limitation we present Holodeck a system that generates 3D environments to match a user-supplied prompt fully automatedly. Holodeck can generate diverse scenes e.g. arcades spas and museums adjust the designs for styles and can capture the semantics of complex queries such as "apartment for a researcher with a cat" and "office of a professor who is a fan of Star Wars". Holod

eck leverages a large language model (i.e. GPT-4) for common sense knowledge abo
ut what the scene might look like and uses a large collection of 3D assets from
Objaverse to populate the scene with diverse objects. To address the challenge o
f positioning objects correctly we prompt GPT-4 to generate spatial relational c
onstraints between objects and then optimize the layout to satisfy those constra
ints. Our large-scale human evaluation shows that annotators prefer Holodeck ove
r manually designed procedural baselines in residential scenes and that Holodeck
 can produce high-quality outputs for diverse scene types. We also demonstrate a
n exciting application of Holodeck in Embodied AI training agents to navigate in
 novel scenes like music rooms and daycares without human-constructed data which
 is a significant step forward in developing general-purpose embodied agents.
****
Distilled Datamodel with Reverse Gradient Matching
Jingwen Ye, Ruonan Yu, Songhua Liu, Xinchao Wang; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11954-1196
3
The proliferation of large-scale AI models trained on extensive datasets has rev
olutionized machine learning. With these models taking on increasingly central r
oles in various applications the need to understand their behavior and enhance i
nterpretability has become paramount. To investigate the impact of changes in tr
aining data on a pre-trained model a common approach is leave-one-out retraining
. This entails systematically altering the training dataset by removing specific
 samples to observe resulting changes within the model. However retraining the m
odel for each altered dataset presents a significant computational challenge giv
en the need to perform this operation for every dataset variation. In this paper
 we introduce an efficient framework for assessing data impact comprising offlin
e training and online evaluation stages. During the offline training phase we ap
proximate the influence of training data on the target model through a distilled
 synset formulated as a reversed gradient matching problem. For online evaluatio
n we expedite the leave-one-out process using the synset which is then utilized
to compute the attribution matrix based on the evaluation objective. Experimenta
l evaluations including training data attribution and assessments of data qualit
y demonstrate that our proposed method achieves comparable model behavior evalua
tion while significantly speeding up the process compared to the direct retraini
ng method.
****
DistriFusion: Distributed Parallel Inference for High-Resolution Diffusion Model
s
Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing
 Jia, Kai Li, Song Han; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2024, pp. 7183-7193
Diffusion models have achieved great success in synthesizing high-quality images
. However generating high-resolution images with diffusion models is still chall
enging due to the enormous computational costs resulting in a prohibitive latenc
y for interactive applications. In this paper we propose DistriFusion to tackle
this problem by leveraging parallelism across multiple GPUs. Our method splits t
he model input into multiple patches and assigns each patch to a GPU. However na
ively implementing such an algorithm breaks the interaction between patches and
loses fidelity while incorporating such an interaction will incur tremendous com
munication overhead. To overcome this dilemma we observe the high similarity bet
ween the input from adjacent diffusion steps and propose Displaced Patch Paralle
lism which takes advantage of the sequential nature of the diffusion process by
reusing the pre-computed feature maps from the previous timestep to provide cont
ext for the current step. Therefore our method supports asynchronous communicati
on which can be pipelined by computation. Extensive experiments show that our me
thod can be applied to recent Stable Diffusion XL with no quality degradation an
d achieve up to a 6.1x speedup on eight NVIDIA A100s compared to one. Our code i
s publicly available at https://github.com/mit-han-lab/distrifuser.
****
Improving the Generalization of Segmentation Foundation Model under Distribution

Shift via Weakly Supervised Adaptation
Haojie Zhang, Yongyi Su, Xun Xu, Kui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23385-23395
The success of large language models has inspired the computer vision community to explore image segmentation foundation model that is able to zero/few-shot generalize through prompt engineering. Segment-Anything (SAM) among others is the state-of-the-art image segmentation foundation model demonstrating strong zero/few-shot generalization. Despite the success recent studies reveal the weakness of SAM under strong distribution shift. In particular SAM performs awkwardly on corrupted natural images camouflaged images medical images etc. Motivated by the observations we aim to develop a self-training based strategy to adapt SAM to target distribution. Given the unique challenges of large source dataset high computation cost and incorrect pseudo label we propose a weakly supervised self-training architecture with anchor regularization and low-rank finetuning to improve the robustness and computation efficiency of adaptation. We validate the effectiveness on 5 types of downstream segmentation tasks including natural clean/corrupted images medical images camouflaged images and robotic images. Our proposed method is task-agnostic in nature and outperforms pre-trained SAM and state-of-the-art domain adaptation methods on almost all downstream tasks with the same testing prompt inputs.
****
Pseudo Label Refinery for Unsupervised Domain Adaptation on Cross-dataset 3D Object Detection
Zhanwei Zhang, Minghao Chen, Shuai Xiao, Liang Peng, Hengjia Li, Binbin Lin, Ping Li, Wenxiao Wang, Boxi Wu, Deng Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15291-15300
Recent self-training techniques have shown notable improvements in unsupervised domain adaptation for 3D object detection (3D UDA). These techniques typically select pseudo labels i.e. 3D boxes to supervise models for the target domain. However this selection process inevitably introduces unreliable 3D boxes in which 3D points cannot be definitively assigned as foreground or background. Previous techniques mitigate this by reweighting these boxes as pseudo labels but these boxes can still poison the training process. To resolve this problem in this paper we propose a novel pseudo label refinery framework. Specifically in the selection process to improve the reliability of pseudo boxes we propose a complementary augmentation strategy. This strategy involves either removing all points within an unreliable box or replacing it with a high-confidence box. Moreover the point numbers of instances in high-beam datasets are considerably higher than those in low-beam datasets also degrading the quality of pseudo labels during the training process. We alleviate this issue by generating additional proposals and aligning RoI features across different domains. Experimental results demonstrate that our method effectively enhances the quality of pseudo labels and consistently surpasses the state-of-the-art methods on six autonomous driving benchmarks. Code will be available at https://github.com/Zhanwei-Z/PERE.
****
Reconstructing Hands in 3D with Transformers
Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, Jitendra Malik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9826-9836
We present an approach that can reconstruct hands in 3D from monocular input. Our approach for Hand Mesh Recovery HaMeR follows a fully transformer-based architecture and can analyze hands with significantly increased accuracy and robustness compared to previous work. The key to HaMeR's success lies in scaling up both the data used for training and the capacity of the deep network for hand reconstruction. For training data we combine multiple datasets that contain 2D or 3D hand annotations. For the deep model we use a large scale Vision Transformer architecture. Our final model consistently outperforms the previous baselines on popular 3D hand pose benchmarks. To further evaluate the effect of our design in non-controlled settings we annotate existing in-the-wild datasets with 2D hand keypoint annotations. On this newly collected dataset of annotations HInt we demonst

rate significant improvements over existing baselines. We will make our code dat a and models publicly available upon publication. We make our code data and mode ls available on the project website: https://geopavlakos.github.io/hamer/.
****
AZ-NAS: Assembling Zero-Cost Proxies for Network Architecture Search
Junghyup Lee, Bumsub Ham; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2024, pp. 5893-5903
Training-free network architecture search (NAS) aims to discover high-performing networks with zero-cost proxies capturing network characteristics related to th e final performance. However network rankings estimated by previous training-fre e NAS methods have shown weak correlations with the performance. To address this issue we propose AZ-NAS a novel approach that leverages the ensemble of various zero-cost proxies to enhance the correlation between a predicted ranking of net works and the ground truth substantially in terms of the performance. To achieve this we introduce four novel zero-cost proxies that are complementary to each o ther analyzing distinct traits of architectures in the views of expressivity pro gressivity trainability and complexity. The proxy scores can be obtained simulta neously within a single forward and backward pass making an overall NAS process highly efficient. In order to integrate the rankings predicted by our proxies ef fectively we introduce a non-linear ranking aggregation method that highlights t he networks highly-ranked consistently across all the proxies. Experimental resu lts conclusively demonstrate the efficacy and efficiency of AZ-NAS outperforming state-of-the-art methods on standard benchmarks all while maintaining a reasona ble runtime cost.
****
Correspondence-Free Non-Rigid Point Set Registration Using Unsupervised Clusteri ng Analysis
Mingyang Zhao, Jingen Jiang, Lei Ma, Shiqing Xin, Gaofeng Meng, Dong-Ming Yan; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21199-21208
This paper presents a novel non-rigid point set registration method that is insp ired by unsupervised clustering analysis. Unlike previous approaches that treat the source and target point sets as separate entities we develop a holistic fram ework where they are formulated as clustering centroids and clustering members s eparately. We then adopt Tikhonov regularization with an ?1-induced Laplacian ke rnel instead of the commonly used Gaussian kernel to ensure smooth and more robu st displacement fields. Our formulation delivers closed-form solutions theoretic al guarantees independence from dimensions and the ability to handle large defor mations. Subsequently we introduce a clustering-improved Nystrom method to effec tively reduce the computational complexity and storage of the Gram matrix to lin ear while providing a rigorous bound for the low rank approximation. Our method achieves high accuracy results across various scenarios and surpasses competitor s by a significant margin particularly on shapes with substantial deformations. Additionally we demonstrate the versatility of our method in challenging tasks s uch as shape transfer and medical registration.
****
Improving Physics-Augmented Continuum Neural Radiance Field-Based Geometry-Agnos tic System Identification with Lagrangian Particle Optimization
Takuhiro Kaneko; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2024, pp. 5470-5480
Geometry-agnostic system identification is a technique for identifying the geome try and physical properties of an object from video sequences without any geomet ric assumptions. Recently physics-augmented continuum neural radiance fields (PA C-NeRF) has demonstrated promising results for this technique by utilizing a hyb rid Eulerian-Lagrangian representation in which the geometry is represented by t he Eulerian grid representations of NeRF the physics is described by a material point method (MPM) and they are connected via Lagrangian particles. However a no table limitation of PAC-NeRF is that its performance is sensitive to the learnin g of the geometry from the first frames owing to its two-step optimization. Firs t the grid representations are optimized with the first frames of video sequence

s and then the physical properties are optimized through video sequences utilizi ng the fixed first-frame grid representations. This limitation can be critical w hen learning of the geometric structure is difficult for example in a few-shot (sparse view) setting. To overcome this limitation we propose Lagrangian particle optimization (LPO) in which the positions and features of particles are optimiz ed through video sequences in Lagrangian space. This method allows for the optim ization of the geometric structure across the entire video sequence within the p hysical constraints imposed by the MPM. The experimental results demonstrate tha t the LPO is useful for geometric correction and physical identification in spar se-view settings.
****
BadCLIP: Trigger-Aware Prompt Learning for Backdoor Attacks on CLIP
Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, Wei Liu; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 24239-24250
Contrastive Vision-Language Pre-training known as CLIP has shown promising effec tiveness in addressing downstream image recognition tasks. However recent works revealed that the CLIP model can be implanted with a downstream-oriented backdoo r. On downstream tasks one victim model performs well on clean samples but predi cts a specific target class whenever a specific trigger is present. For injectin g a backdoor existing attacks depend on a large amount of additional data to mal iciously fine-tune the entire pre-trained CLIP model which makes them inapplicab le to data-limited scenarios. In this work motivated by the recent success of le arnable prompts we address this problem by injecting a backdoor into the CLIP mo del in the prompt learning stage. Our method named BadCLIP is built on a novel a nd effective mechanism in backdoor attacks on CLIP i.e. influencing both the ima ge and text encoders with the trigger. It consists of a learnable trigger applie d to images and a trigger-aware context generator such that the trigger can chan ge text features via trigger-aware prompts resulting in a powerful and generaliz able attack. Extensive experiments conducted on 11 datasets verify that the clea n accuracy of BadCLIP is similar to those of advanced prompt learning methods an d the attack success rate is higher than 99% in most cases. BadCLIP is also gene ralizable to unseen classes and shows a strong generalization capability under c ross-dataset and cross-domain settings. The code is available at https://github. com/jiawangbai/BadCLIP.
****
Beyond Image Super-Resolution for Image Recognition with Task-Driven Perceptual Loss
Jaeha Kim, Junghun Oh, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2651-2661
In real-world scenarios image recognition tasks such as semantic segmentation an d object detection often pose greater challenges due to the lack of information available within low-resolution (LR) content. Image super-resolution (SR) is one of the promising solutions for addressing the challenges. However due to the il l-posed property of SR it is challenging for typical SR methods to restore task-relevant high-frequency contents which may dilute the advantage of utilizing the SR method. Therefore in this paper we propose Super-Resolution for Image Recogn ition (SR4IR) that effectively guides the generation of SR images beneficial to achieving satisfactory image recognition performance when processing LR images. The critical component of our SR4IR is the task-driven perceptual (TDP) loss tha t enables the SR network to acquire task-specific knowledge from a network tailo red for a specific task. Moreover we propose a cross-quality patch mix and an al ternate training framework that significantly enhances the efficacy of the TDP l oss by addressing potential problems when employing the TDP loss. Through extens ive experiments we demonstrate that our SR4IR achieves outstanding task performa nce by generating SR images useful for a specific image recognition task includi ng semantic segmentation object detection and image classification. The implemen tation code is available at https://github.com/JaehaKim97/SR4IR.
****
PELA: Learning Parameter-Efficient Models with Low-Rank Approximation

Yangyang Guo, Guangzhi Wang, Mohan Kankanhalli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15699-15709
Applying a pre-trained large model to downstream tasks is prohibitive under resource-constrained conditions. Recent dominant approaches for addressing efficiency issues involve adding a few learnable parameters to the fixed backbone model. This strategy however leads to more challenges in loading large models for downstream fine-tuning with limited resources. In this paper we propose a novel method for increasing the parameter efficiency of pre-trained models by introducing an intermediate pre-training stage. To this end we first employ low-rank approximation to compress the original large model and then devise a feature distillation module and a weight perturbation regularization module. These modules are specifically designed to enhance the low-rank model. In particular we update only the low-rank model while freezing the backbone parameters during pre-training. This allows for direct and efficient utilization of the low-rank model for downstream fine-tuning tasks. The proposed method achieves both efficiencies in terms of required parameters and computation time while maintaining comparable results with minimal modifications to the backbone architecture. Specifically when applied to three vision-only and one vision-language Transformer models our approach often demonstrates a merely 0.6-point decrease in performance while reducing the original parameter size by 1/3 to 2/3.
****
XCube: Large-Scale 3D Generative Modeling using Sparse Voxel Hierarchies
Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, Francis Williams; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4209-4219
We present XCube a novel generative model for high-resolution sparse 3D voxel grids with arbitrary attributes. Our model can generate millions of voxels with a finest effective resolution of up to 1024^3 in a feed-forward fashion without time-consuming test-time optimization. To achieve this we employ a hierarchical voxel latent diffusion model which generates progressively higher resolution grids in a coarse-to-fine manner using a custom framework built on the highly efficient VDB data structure. Apart from generating high-resolution objects we demonstrate the effectiveness of XCube on large outdoor scenes at scales of 100m x 100m with a voxel size as small as 10cm. We observe clear qualitative and quantitative improvements over past approaches. In addition to unconditional generation we show that our model can be used to solve a variety of tasks such as user-guided editing scene completion from a single scan and text-to-3D.
****
PixelRNN: In-pixel Recurrent Neural Networks for End-to-end-optimized Perception with Neural Sensors
Haley M. So, Laurie Bose, Piotr Dudek, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25233-25244
Conventional image sensors digitize high-resolution images at fast frame rates producing a large amount of data that needs to be transmitted off the sensor for further processing. This is challenging for perception systems operating on edge devices because communication is power inefficient and induces latency. Fueled by innovations in stacked image sensor fabrication emerging sensor--processors offer programmability and processing capabilities directly on the sensor. We exploit these capabilities by developing an efficient recurrent neural network architecture PixelRNN that encodes spatio-temporal features on the sensor using purely binary operations. PixelRNN reduces the amount of data to be transmitted off the sensor by factors up to 256 compared to the raw sensor data while offering competitive accuracy for hand gesture recognition and lip reading tasks. We experimentally validate PixelRNN using a prototype implementation on the SCAMP-5 sensor--processor platform.
****
Reconstruction-free Cascaded Adaptive Compressive Sensing
Chenxi Qiu, Tao Yue, Xuemei Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2620-2630

Scene-aware Adaptive Compressive Sensing (ACS) has constituted a persistent pursuit holding substantial promise for the enhancement of Compressive Sensing (CS) performance. Cascaded ACS furnishes a proficient multi-stage framework for adaptively allocating the CS sampling based on previous CS measurements. However reconstruction is commonly required for analyzing and steering the successive CS sampling which bottlenecks the ACS speed and impedes the practical application in time-sensitive scenarios. Addressing this challenge we propose a reconstruction-free cascaded ACS method which requires NO reconstruction during the adaptive sampling process. A lightweight Score Network (ScoreNet) is proposed to directly determine the ACS allocation with previous CS measurements and a differentiable adaptive sampling module is proposed for end-to-end training. For image reconstruction we propose a Multi-Grid Spatial-Attention Network (MGSANet) that could facilitate efficient multi-stage training and inferencing. By introducing the reconstruction-fidelity supervision outside the loop of the multi-stage sampling process ACS can be efficiently optimized and achieve high imaging fidelity. The effectiveness of the proposed method is demonstrated with extensive quantitative and qualitative experiments compared with the state-of-the-art CS algorithms.

****

Auto-Train-Once: Controller Network Guided Automatic Network Pruning from Scratch

Xidong Wu, Shangqian Gao, Zeyu Zhang, Zhenzhen Li, Runxue Bao, Yanfu Zhang, Xiaoqian Wang, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16163-16173

Current techniques for deep neural network (DNN) pruning often involve intricate multi-step processes that require domain-specific expertise making their widespread adoption challenging. To address the limitation the Only-Train-Once (OTO) and OTOv2 are proposed to eliminate the need for additional fine-tuning steps by directly training and compressing a general DNN from scratch. Nevertheless the static design of optimizers (in OTO) can lead to convergence issues of local optima. In this paper we proposed the Auto-Train-Once (ATO) an innovative network pruning algorithm designed to automatically reduce the computational and storage costs of DNNs. During the model training phase our approach not only trains the target model but also leverages a controller network as an architecture generator to guide the learning of target model weights. Furthermore we developed a novel stochastic gradient algorithm that enhances the coordination between model training and controller network training thereby improving pruning performance. We provide a comprehensive convergence analysis as well as extensive experiments and the results show that our approach achieves state-of-the-art performance across various model architectures (including ResNet18 ResNet34 ResNet50 ResNet56 and MobileNetv2) on standard benchmark datasets (CIFAR-10 CIFAR-100 and ImageNet).

****

Constructing and Exploring Intermediate Domains in Mixed Domain Semi-supervised Medical Image Segmentation

Qinghe Ma, Jian Zhang, Lei Qi, Qian Yu, Yinghuan Shi, Yang Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11642-11651

Both limited annotation and domain shift are prevalent challenges in medical image segmentation. Traditional semi-supervised segmentation and unsupervised domain adaptation methods address one of these issues separately. However the coexistence of limited annotation and domain shift is quite common which motivates us to introduce a novel and challenging scenario: Mixed Domain Semi-supervised medical image Segmentation (MiDSS). In this scenario we handle data from multiple medical centers with limited annotations available for a single domain and a large amount of unlabeled data from multiple domains. We found that the key to solving the problem lies in how to generate reliable pseudo labels for the unlabeled data in the presence of domain shift with labeled data. To tackle this issue we employ Unified Copy-Paste (UCP) between images to construct intermediate domains facilitating the knowledge transfer from the domain of labeled data to the domains of unlabeled data. To fully utilize the information within the intermediate domain we propose a symmetric Guidance training strategy (SymGD) which additionall

y offers direct guidance to unlabeled data by merging pseudo labels from interme
diate samples. Subsequently we introduce a Training Process aware Random Amplitu
de MixUp (TP-RAM) to progressively incorporate style-transition components into
intermediate samples. Compared with existing state-of-the-art approaches our met
hod achieves a notable 13.57% improvement in Dice score on Prostate dataset as d
emonstrated on three public datasets. Our code is available at https://github.co
m/MQinghe/MiDSS
****
DUSt3R: Geometric 3D Vision Made Easy
Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, Jerome Revaud; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 20697-20709
Multi-view stereo reconstruction (MVS) in the wild requires to first estimate th
e camera intrinsic and extrinsic parameters. These are usually tedious and cumbe
rsome to obtain yet they are mandatory to triangulate corresponding pixels in 3D
 space which is at the core of all best performing MVS algorithms. In this work
we take an opposite stance and introduce DUSt3R a radically novel paradigm for D
ense and Unconstrained Stereo 3D Reconstruction of arbitrary image collections o
perating without prior information about camera calibration nor viewpoint poses.
 We cast the pairwise reconstruction problem as a regression of pointmaps relaxi
ng the hard constraints of usual projective camera models. We show that this for
mulation smoothly unifies the monocular and binocular reconstruction cases. In t
he case where more than two images are provided we further propose a simple yet
effective global alignment strategy that expresses all pairwise pointmaps in a c
ommon reference frame. We base our network architecture on standard Transformer
encoders and decoders allowing us to leverage powerful pretrained models. Our fo
rmulation directly provides a 3D model of the scene as well as depth information
 but interestingly we can seamlessly recover from it pixel matches focal lengths
 relative and absolute cameras. Extensive experiments on all these tasks showcas
e how DUSt3R effectively unifies various 3D vision tasks setting new performance
 records on monocular & multi-view depth estimation as well as relative pose est
imation. In summary DUSt3R makes many geometric 3D vision tasks easy. Code and m
odels at https://github.com/naver/dust3r
****
From Isolated Islands to Pangea: Unifying Semantic Space for Human Action Unders
tanding
Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Zehao Wang, Yiming Dou, Yikun Ji, Junyi Zh
ang, Yixing Li, Xudong Lu, Jingru Tan, Cewu Lu; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16582-16592
Action understanding matters for intelligent agents and has attracted long-term
attention. It can be formed as the mapping from the action physical space to the
 semantic space. Typically researchers built action datasets according to idiosy
ncratic choices to define classes and push the envelope of benchmarks respective
ly. Thus datasets are incompatible with each other like "Isolated Islands" due t
o semantic gaps and various class granularities e.g. do housework in dataset A a
nd wash plate in dataset B. We argue that a more principled semantic space is an
 urgent need to concentrate the community efforts and enable us to use all datas
ets together to pursue generalizable action learning. To this end we design a st
ructured action semantic space in view of verb taxonomy hierarchy and covering m
assive actions. By aligning the classes of previous datasets to our semantic spa
ce we gather (image/video/skeleton/MoCap) datasets into a unified database in a
unified label system i.e. bridging "isolated islands" into a "Pangea". According
ly we propose a novel model mapping from the physical space to semantic space to
 fully use Pangea. In extensive experiments our new system shows significant sup
eriority especially in transfer learning. Our code and data will be made public
at https://mvig-rhos.com/pangea.
****
Bootstrapping Autonomous Driving Radars with Self-Supervised Learning
Yiduo Hao, Sohrab Madani, Junfeng Guan, Mohammed Alloulah, Saurabh Gupta, Haitha
m Hassanieh; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte

rn Recognition (CVPR), 2024, pp. 15012-15023
The perception of autonomous vehicles using radars has attracted increased resea rch interest due its ability to operate in fog and bad weather. However training radar models is hindered by the cost and difficulty of annotating large-scale r adar data. To overcome this bottleneck we propose a self-supervised learning fra mework to leverage the large amount of unlabeled radar data to pre-train radar-o nly embeddings for self-driving perception tasks. The proposed method combines r adar-to-radar and radar-to-vision contrastive losses to learn a general represen tation from unlabeled radar heatmaps paired with their corresponding camera imag es. When used for downstream object detection we demonstrate that the proposed s elf-supervision framework can improve the accuracy of state-of-the-art supervise d baselines by 5.8% in mAP. Code is available at https://github.com/yiduohao/Rad ical.
****
Robust Distillation via Untargeted and Targeted Intermediate Adversarial Samples
Junhao Dong, Piotr Koniusz, Junxi Chen, Z. Jane Wang, Yew-Soon Ong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 24, pp. 28432-28442
Adversarially robust knowledge distillation aims to compress large-scale models into lightweight models while preserving adversarial robustness and natural perf ormance on a given dataset. Existing methods typically align probability distrib utions of natural and adversarial samples between teacher and student models but they overlook intermediate adversarial samples along the "adversarial path" for med by the multi-step gradient ascent of a sample towards the decision boundary. Such paths capture rich information about the decision boundary. In this paper we propose a novel adversarially robust knowledge distillation approach by incor porating such adversarial paths into the alignment process. Recognizing the dive rse impacts of intermediate adversarial samples (ranging from benign to noisy) w e propose an adaptive weighting strategy to selectively emphasize informative ad versarial samples thus ensuring efficient utilization of lightweight model capac ity. Moreover we propose a dual-branch mechanism exploiting two following insigh ts: (i) complementary dynamics of adversarial paths obtained by targeted and unt argeted adversarial learning and (ii) inherent differences between the gradient ascent path from class c_i towards the nearest class boundary and the gradient d escent path from a specific class c_j towards the decision region of c_i (i \neq j). Comprehensive experiments demonstrate the effectiveness of our method on li ghtweight models under various settings.
****
USE: Universal Segment Embeddings for Open-Vocabulary Image Segmentation
Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li , Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, Liu Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp . 4187-4196
The open-vocabulary image segmentation task involves partitioning images into se mantically meaningful segments and classifying them with flexible text-defined c ategories. The recent vision-based foundation models such as the Segment Anythin g Model (SAM) have shown superior performance in generating class-agnostic image segments. The main challenge in open-vocabulary image segmentation now lies in accurately classifying these segments into text-defined categories. In this pape r we introduce the Universal Segment Embedding (USE) framework to address this c hallenge. This framework is comprised of two key components: 1) a data pipeline designed to efficiently curate a large amount of segment-text pairs at various g ranularities and 2) a universal segment embedding model that enables precise seg ment classification into a vast range of text-defined categories. The USE model can not only help open-vocabulary image segmentation but also facilitate other d ownstream tasks (e.g. querying and ranking). Through comprehensive experimental studies on semantic segmentation and part segmentation benchmarks we demonstrate that the USE framework outperforms state-of-the-art open-vocabulary segmentatio n methods.
****

Functional Diffusion
Biao Zhang, Peter Wonka; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 4723-4732
We propose functional diffusion a generative diffusion model focused on infinite
-dimensional function data samples. In contrast to previous work functional diff
usion works on samples that are represented by functions with a continuous domai
n. Functional diffusion can be seen as an extension of classical diffusion model
s to an infinite-dimensional domain. Functional diffusion is very versatile as i
mages videos audio 3D shapes deformations etc. can be handled by the same framew
ork with minimal changes. In addition functional diffusion is especially suited
for irregular data or data defined in non-standard domains. In our work we deriv
e the necessary foundations for functional diffusion and propose a first impleme
ntation based on the transformer architecture. We show generative results on com
plicated signed distance functions and deformation functions defined on 3D surfa
ces.
****
Soften to Defend: Towards Adversarial Robustness via Self-Guided Label Refinemen
t
Zhuorong Li, Daiwei Yu, Lina Wei, Canghong Jin, Yun Zhang, Sixian Chan; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2024, pp. 24776-24785
Adversarial training (AT) is currently one of the most effective ways to obtain
the robustness of deep neural networks against adversarial attacks. However most
 AT methods suffer from robust overfitting i.e. a significant generalization gap
 in adversarial robustness between the training and testing curves. In this pape
r we first identify a connection between robust overfitting and the excessive me
morization of noisy labels in AT from a view of gradient norm. As such label noi
se is mainly caused by a distribution mismatch and improper label assignments we
 are motivated to propose a label refinement approach for AT. Specifically our S
elf-Guided Label Refinement first self-refines a more accurate and informative l
abel distribution from over-confident hard labels and then it calibrates the tra
ining by dynamically incorporating knowledge from self-distilled models into the
 current model and thus requiring no external teachers. Empirical results demons
trate that our method can simultaneously boost the standard accuracy and robust
performance across multiple benchmark datasets attack types and architectures. I
n addition we also provide a set of analyses from the perspectives of informatio
n theory to dive into our method and suggest the importance of soft labels for r
obust generalization.
****
Weakly Supervised Monocular 3D Detection with a Single-View Image
Xueying Jiang, Sheng Jin, Lewei Lu, Xiaoqin Zhang, Shijian Lu; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 10508-10518
Monocular 3D detection (M3D) aims for precise 3D object localization from a sing
le-view image which usually involves labor-intensive annotation of 3D detection
boxes. Weakly supervised M3D has recently been studied to obviate the 3D annotat
ion process by leveraging many existing 2D annotations but it often requires ext
ra training data such as LiDAR point clouds or multi-view images which greatly d
egrades its applicability and usability in various applications. We propose SKD-
WM3D a weakly supervised monocular 3D detection framework that exploits depth in
formation to achieve M3D with a single-view image exclusively without any 3D ann
otations or other training data. One key design in SKD-WM3D is a self-knowledge
distillation framework which transforms image features into 3D-like representati
ons by fusing depth information and effectively mitigates the inherent depth amb
iguity in monocular scenarios with little computational overhead in inference. I
n addition we design an uncertainty-aware distillation loss and a gradient-targe
ted transfer modulation strategy which facilitate knowledge acquisition and know
ledge transfer respectively. Extensive experiments show that SKD-WM3D surpasses
the state-of-the-art clearly and is even on par with many fully supervised metho
ds.

****
Pose-Guided Self-Training with Two-Stage Clustering for Unsupervised Landmark Di
scovery
Siddharth Tourani, Ahmed Alwheibi, Arif Mahmood, Muhammad Haris Khan; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 23041-23051
Unsupervised landmarks discovery (ULD) for an object category is a challenging c
omputer vision problem. In pursuit of developing a robust ULD framework we explo
re the potential of a recent paradigm of self-supervised learning algorithms kno
wn as diffusion models. Some recent works have shown that these models implicitl
y contain important correspondence cues. Towards harnessing the potential of dif
fusion models for ULD task we make the following core contributions. First we pr
opose a ZeroShot ULD baseline based on simple clustering of random pixel locatio
ns with nearest neighbour matching. It delivers better results than the existing
 ULD methods. Second motivated by the ZeroShot performance we develop a ULD algo
rithm based on diffusion features using self-training and clustering which also
outperforms prior methods by notable margins. Third we introduce a new proxy tas
k based on generating latent pose codes and also propose a two-stage clustering
mechanism to facilitate effective pseudo-labeling resulting in a significant per
formance improvement. Overall our approach consistently outperforms state-of-the
-art methods on four challenging benchmarks AFLW MAFL CatHeads and LS3D by signi
ficant margins.
****
Learning from Synthetic Human Group Activities
Che-Jui Chang, Danrui Li, Deep Patel, Parth Goel, Honglu Zhou, Seonghyeon Moon,
Samuel S. Sohn, Sejong Yoon, Vladimir Pavlovic, Mubbasir Kapadia; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 21922-21932
The study of complex human interactions and group activities has become a focal
point in human-centric computer vision. However progress in related tasks is oft
en hindered by the challenges of obtaining large-scale labeled datasets from rea
l-world scenarios. To address the limitation we introduce M3Act a synthetic data
 generator for multi-view multi-group multi-person human atomic actions and grou
p activities. Powered by Unity Engine M3Act features multiple semantic groups hi
ghly diverse and photorealistic images and a comprehensive set of annotations wh
ich facilitates the learning of human-centered tasks across single-person multi-
person and multi-group conditions. We demonstrate the advantages of M3Act across
 three core experiments. The results suggest our synthetic dataset can significa
ntly improve the performance of several downstream methods and replace real-worl
d datasets to reduce cost. Notably M3Act improves the state-of-the-art MOTRv2 on
 DanceTrack dataset leading to a hop on the leaderboard from 10th to 2nd place.
Moreover M3Act opens new research for controllable 3D group activity generation.
 We define multiple metrics and propose a competitive baseline for the novel tas
k. Our code and data are available at our project page: http://cjerry1243.github
.io/M3Act.
****
Blind Image Quality Assessment Based on Geometric Order Learning
Nyeong-Ho Shin, Seon-Ho Lee, Chang-Su Kim; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12799-12808
A novel approach to blind image quality assessment called quality comparison net
work (QCN) is proposed in this paper which sorts the feature vectors of input im
ages according to their quality scores in an embedding space. QCN employs compar
ison transformers (CTs) and score pivots which act as the centroids of feature v
ectors of similar-quality images. Each CT updates the score pivots and the featu
re vectors of input images based on their ordered correlation. To this end we ad
opt four loss functions. Then we estimate the quality score of a test image by s
earching the nearest score pivot to its feature vector in the embedding space. E
xtensive experiments show that the proposed QCN algorithm yields excellent image
 quality assessment performances on various datasets. Furthermore QCN achieves g
reat performances in cross-dataset evaluation demonstrating its superb generaliz

ation capability. The source codes are available at https://github.com/nhshin-mc
l/QCN.
****
Text Grouping Adapter: Adapting Pre-trained Text Detector for Layout Analysis
Tianci Bi, Xiaoyi Zhang, Zhizheng Zhang, Wenxuan Xie, Cuiling Lan, Yan Lu, Nanni
ng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2024, pp. 28150-28159
Significant progress has been made in scene text detection models since the rise
 of deep learning but scene text layout analysis which aims to group detected te
xt instances as paragraphs has not kept pace. Previous works either treated text
 detection and grouping using separate models or train a model from scratch whil
e using a unified one. All of them have not yet made full use of the already wel
l-trained text detectors and easily obtainable detection datasets. In this paper
 we present Text Grouping Adapter (TGA) a module that can enable the utilization
 of various pre-trained text detectors to learn layout analysis allowing us to a
dopt a well-trained text detector right off the shelf or just fine-tune it effic
iently. Designed to be compatible with various text detector architectures TGA t
akes detected text regions and image features as universal inputs to assemble te
xt instance features. To capture broader contextual information for layout analy
sis we propose to predict text group masks from text instance features by one-to
-many assignment. Our comprehensive experiments demonstrate that even with froze
n pre-trained models incorporating our TGA into various pre-trained text detecto
rs and text spotters can achieve superior layout analysis performance simultaneo
usly inheriting generalized text detection ability from pre-training. In the cas
e of full parameter fine-tuning we can further improve layout analysis performan
ce.
****
Generalizable Whole Slide Image Classification with Fine-Grained Visual-Semantic
 Interaction
Hao Li, Ying Chen, Yifei Chen, Rongshan Yu, Wenxian Yang, Liansheng Wang, Bowen
Ding, Yuchen Han; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2024, pp. 11398-11407
Whole Slide Image (WSI) classification is often formulated as a Multiple Instanc
e Learning (MIL) problem. Recently Vision-Language Models (VLMs) have demonstrat
ed remarkable performance in WSI classification. However existing methods levera
ge coarse-grained pathogenetic descriptions for visual representation supervisio
n which are insufficient to capture the complex visual appearance of pathogeneti
c images hindering the generalizability of models on diverse downstream tasks. A
dditionally processing high-resolution WSIs can be computationally expensive. In
 this paper we propose a novel "Fine-grained Visual-Semantic Interaction" (FiVE)
 framework for WSI classification. It is designed to enhance the model's general
izability by leveraging the interaction between localized visual patterns and fi
ne-grained pathological semantics. Specifically with meticulously designed queri
es we start by utilizing a large language model to extract fine-grained patholog
ical descriptions from various non-standardized raw reports. The output descript
ions are then reconstructed into fine-grained labels used for training. By intro
ducing a Task-specific Fine-grained Semantics (TFS) module we enable prompts to
capture crucial visual information in WSIs which enhances representation learnin
g and augments generalization capabilities significantly. Furthermore given that
 pathological visual patterns are redundantly distributed across tissue slices w
e sample a subset of visual instances during training. Our method demonstrates r
obust generalizability and strong transferability dominantly outperforming the c
ounterparts on the TCGA Lung Cancer dataset with at least 9.19% higher accuracy
in few-shot experiments. The code is available at: https://github.com/ls1rius/WS
I_FiVE.
****
THRONE: An Object-based Hallucination Benchmark for the Free-form Generations of
 Large Vision-Language Models
Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. T
aylor, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR), 2024, pp. 27228-27238
Mitigating hallucinations in large vision-language models (LVLMs) remains an ope
n problem. Recent benchmarks do not address hallucinations in open-ended free-fo
rm responses which we term "Type I hallucinations". Instead they focus on halluc
inations responding to very specific question formats---typically a multiple-cho
ice response regarding a particular object or attribute---which we term "Type II
 hallucinations". Additionally such benchmarks often require external API calls
to models which are subject to change. In practice we observe that a reduction i
n Type II hallucinations does not lead to a reduction in Type I hallucinations b
ut rather that the two forms of hallucinations are often anti-correlated. To add
ress this we propose THRONE a novel object-based automatic framework for quantit
atively evaluating Type I hallucinations in LVLM free-form outputs. We use publi
c language models (LMs) to identify hallucinations in LVLM responses and compute
 informative metrics. By evaluating a large selection of recent LVLMs using publ
ic datasets we show that an improvement in existing metrics do not lead to a red
uction in Type I hallucinations and that established benchmarks for measuring Ty
pe I hallucinations are incomplete. Finally we provide a simple and effective da
ta augmentation method to reduce Type I and Type II hallucinations as a strong b
aseline.
****
Wired Perspectives: Multi-View Wire Art Embraces Generative AI
Zhiyu Qu, Lan Yang, Honggang Zhang, Tao Xiang, Kaiyue Pang, Yi-Zhe Song; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 6149-6158
Creating multi-view wire art (MVWA) a static 3D sculpture with diverse interpret
ations from different viewpoints is a complex task even for skilled artists. In
response we present DreamWire an AI system enabling everyone to craft MVWA easil
y. Users express their vision through text prompts or scribbles freeing them fro
m intricate 3D wire organisation. Our approach synergises 3D Bezier curves Prim'
s algorithm and knowledge distillation from diffusion models or their variants (
e.g. ControlNet). This blend enables the system to represent 3D wire art ensurin
g spatial continuity and overcoming data scarcity. Extensive evaluation and anal
ysis are conducted to shed insight on the inner workings of the proposed system
including the trade-off between connectivity and visual aesthetics.
****
LUWA Dataset: Learning Lithic Use-Wear Analysis on Microscopic Images
Jing Zhang, Irving Fang, Hao Wu, Akshat Kaushik, Alice Rodriguez, Hanwen Zhao, J
uexiao Zhang, Zhuo Zheng, Radu Iovita, Chen Feng; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22563-2257
3
Lithic Use-Wear Analysis (LUWA) using microscopic images is an underexplored vis
ion-for-science research area. It seeks to distinguish the worked material which
 is critical for understanding archaeological artifacts material interactions to
ol functionalities and dental records. However this challenging task goes beyond
 the well-studied image classification problem for common objects. It is affecte
d by many confounders owing to the complex wear mechanism and microscopic imagin
g which makes it difficult even for human experts to identify the worked materia
l successfully. In this paper we investigate the following three questions on th
is unique vision task for the first time:(i) How well can state-of-the-art pre-t
rained models (like DINOv2) generalize to the rarely seen domain? (ii) How can f
ew-shot learning be exploited for scarce microscopic images? (iii) How do the am
biguous magnification and sensing modality influence the classification accuracy
? To study these we collaborated with archaeologists and built the first open-so
urce and the largest LUWA dataset containing 23130 microscopic images with diffe
rent magnifications and sensing modalities. Extensive experiments show that exis
ting pre-trained models notably outperform human experts but still leave a large
 gap for improvements. Most importantly the LUWA dataset provides an underexplor
ed opportunity for vision and learning communities and complements existing imag
e classification problems on common objects.
****

Generalizing 6-DoF Grasp Detection via Domain Prior Knowledge
Haoxiang Ma, Modi Shi, Boyang Gao, Di Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18102-18111
We focus on the generalization ability of the 6-DoF grasp detection method in this paper. While learning-based grasp detection methods can predict grasp poses for unseen objects using the grasp distribution learned from the training set they often exhibit a significant performance drop when encountering objects with diverse shapes and structures. To enhance the grasp detection methods' generalization ability we incorporate domain prior knowledge of robotic grasping enabling better adaptation to objects with significant shape and structure differences. More specifically we employ the physical constraint regularization during the training phase to guide the model towards predicting grasps that comply with the physical rule on grasping. For the unstable grasp poses predicted on novel objects we design a contact-score joint optimization using the projection contact map to refine these poses in cluttered scenarios. Extensive experiments conducted on the GraspNet-1billion benchmark demonstrate a substantial performance gain on the novel object set and the real-world grasping experiments also demonstrate the effectiveness of our generalizing 6-DoF grasp detection method.
****
The Audio-Visual Conversational Graph: From an Egocentric-Exocentric Perspective
Wenqi Jia, Miao Liu, Hao Jiang, Ishwarya Ananthabhotla, James M. Rehg, Vamsi Krishna Ithapu, Ruohan Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26396-26405
In recent years the thriving development of research related to egocentric videos has provided a unique perspective for the study of conversational interactions where both visual and audio signals play a crucial role. While most prior work focus on learning about behaviors that directly involve the camera wearer we introduce the Ego-Exocentric Conversational Graph Prediction problem marking the first attempt to infer exocentric conversational interactions from egocentric videos. We propose a unified multi-modal framework---Audio-Visual Conversational Attention (AV-CONV) for the joint prediction of conversation behaviors---speaking and listening---for both the camera wearer as well as all other social partners present in the egocentric video. Specifically we adopt the self-attention mechanism to model the representations across-time across-subjects and across-modalities. To validate our method we conduct experiments on a challenging egocentric video dataset that includes multi-speaker and multi-conversation scenarios. Our results demonstrate the superior performance of our method compared to a series of baselines. We also present detailed ablation studies to assess the contribution of each component in our model. Check our \href https://vjwq.github.io/AV-CONV/ Project Page .
****
Byzantine-robust Decentralized Federated Learning via Dual-domain Clustering and Trust Bootstrapping
Peng Sun, Xinyang Liu, Zhibo Wang, Bo Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24756-24765
Decentralized federated learning (DFL) facilitates collaborative model training across multiple connected clients without a central coordination server thereby avoiding the single point of failure in traditional centralized federated learning (CFL). However DFL exhibits heightened susceptibility to Byzantine attacks owing to the lack of a responsible central server. Furthermore a benign client in DFL may be dominated by Byzantine clients (more than half of its neighbors are malicious) posing significant challenges for robust model training. In this work we propose DFL-Dual a novel Byzantine-robust DFL method through dual-domain client clustering and trust bootstrapping. Specifically we first propose to leverage both data-domain and model-domain distance metrics to identify client discrepancies. Then we design a trust evaluation mechanism centered on benign clients which enables them to evaluate their neighbors. Building upon the dual-domain distance metric and trust evaluation mechanism we further develop a two-stage clustering and trust bootstrapping technique to exclude Byzantine clients from local model aggregation. We extensively evaluate the proposed DFL-Dual method through ri

gorous experimentation demonstrating its remarkable performance superiority over existing robust CFL and DFL schemes.
****
Leveraging Camera Triplets for Efficient and Accurate Structure-from-Motion
Lalit Manam, Venu Madhav Govindu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4959-4968
In Structure-from-Motion (SfM) the underlying viewgraphs of unordered image collections generally have a highly redundant set of edges that can be sparsified for efficiency without significant loss of reconstruction quality. Often there are also false edges due to incorrect image retrieval and repeated structures (symmetries) that give rise to ghosting and superimposed reconstruction artifacts. We present a unified method to simultaneously sparsify the viewgraph and remove false edges. We propose a scoring mechanism based on camera triplets that identifies edge redundancy as well as false edges. Our edge selection is formulated as an optimization problem which can be provably solved using a simple thresholding scheme. This results in a highly efficient algorithm which can be incorporated as a pre-processing step into any SfM pipeline making it practically usable. We demonstrate the utility of our method on generic and ambiguous datasets that cover the range of small medium and large-scale datasets all with different statistical properties. Sparsification of generic datasets using our method significantly reduces reconstruction time while maintaining the accuracy of the reconstructions as well as removing ghosting artifacts. For ambiguous datasets our method removes false edges thereby avoiding incorrect superimposed reconstructions.
****
SimDA: Simple Diffusion Adapter for Efficient Video Generation
Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7827-7839
The recent wave of AI-generated content has witnessed the great development and success of Text-to-Image (T2I) technologies. By contrast Text-to-Video (T2V) still falls short of expectations though attracting increasing interest. Existing works either train from scratch or adapt large T2I model to videos both of which are computation and resource expensive. In this work we propose a Simple Diffusion Adapter (SimDA) that fine-tunes only 24M out of 1.1B parameters of a strong T2I model adapting it to video generation in a parameter-efficient way. In particular we turn the T2I model for T2V by designing light-weight spatial and temporal adapters for transfer learning. Besides we change the original spatial attention to the proposed Latent-Shift Attention (LSA) for temporal consistency. With a similar model architecture we further train a video super-resolution model to generate high-definition (1024 x 1024) videos. In addition to T2V generation in the wild SimDA could also be utilized in one-shot video editing with only 2 minutes tuning. Doing so our method could minimize the training effort with extremely few tunable parameters for model adaptation.
****
Multi-view Aggregation Network for Dichotomous Image Segmentation
Qian Yu, Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3921-3930
Dichotomous Image Segmentation (DIS) has recently emerged towards high-precision object segmentation from high-resolution natural images. When designing an effective DIS model the main challenge is how to balance the semantic dispersion of high-resolution targets in the small receptive field and the loss of high-precision details in the large receptive field. Existing methods rely on tedious multiple encoder-decoder streams and stages to gradually complete the global localization and local refinement. Human visual system captures regions of interest by observing them from multiple views. Inspired by it we model DIS as a multi-view object perception problem and provide a parsimonious multi-view aggregation network (MVANet) which unifies the feature fusion of the distant view and close-up view into a single stream with one encoder-decoder structure. With the help of the proposed multi-view complementary localization and refinement modules our appro

ach established long-range profound visual interactions across multiple views allowing the features of the detailed close-up view to focus on highly slender structures. Experiments on the popular DIS-5K dataset show that our MVANet significantly outperforms state-of-the-art methods in both accuracy and speed. The source code and datasets will be publicly available at \href https://github.com/qianyu-dlut/MVANet MVANet .
****
A Recipe for Scaling up Text-to-Video Generation with Text-free Videos
Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, Nong Sang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6572-6582
Diffusion-based text-to-video generation has witnessed impressive progress in the past year yet still falls behind text-to-image generation. One of the key reasons is the limited scale of publicly available data (e.g. 10M video-text pairs in WebVid10M vs. 5B image-text pairs in LAION) considering the high cost of video captioning. Instead it could be far easier to collect unlabeled clips from video platforms like YouTube. Motivated by this we come up with a novel text-to-video generation framework termed TF-T2V which can directly learn with text-free videos. The rationale behind is to separate the process of text decoding from that of temporal modeling. To this end we employ a content branch and a motion branch which are jointly optimized with weights shared. Following such a pipeline we study the effect of doubling the scale of training set (i.e. video-only WebVid10M) with some randomly collected text-free videos and are encouraged to observe the performance improvement (FID from 9.67 to 8.19 and FVD from 484 to 441) demonstrating the scalability of our approach. We also find that our model could enjoy sustainable performance gain (FID from 8.19 to 7.64 and FVD from 441 to 366) after reintroducing some text labels for training. Finally we validate the effectiveness and generalizability of our ideology on both native text-to-video generation and compositional video synthesis paradigms. Code and models will be publicly available at here.
****
Molecular Data Programming: Towards Molecule Pseudo-labeling with Systematic Weak Supervision
Xin Juan, Kaixiong Zhou, Ninghao Liu, Tianlong Chen, Xin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 308-318
The premise for the great advancement of molecular machine learning is dependent on a considerable amount of labeled data. In many real-world scenarios the labeled molecules are limited in quantity or laborious to derive. Recent pseudo-labeling methods are usually designed based on a single domain knowledge thereby failing to understand the comprehensive molecular configurations and limiting their adaptability to generalize across diverse biochemical context. To this end we introduce an innovative paradigm for dealing with the molecule pseudo-labeling named as Molecular Data Programming (MDP). In particular we adopt systematic supervision sources via crafting multiple graph labeling functions which covers various molecular structural knowledge of graph kernels molecular fingerprints and topological features. Each of them creates an uncertain and biased labels for the unlabeled molecules. To address the decision conflicts among the diverse pseudo-labels we design a label synchronizer to differentiably model confidences and correlations between the labeling functions which yields probabilistic molecular labels to adapt for specific applications. These probabilistic molecular labels are used to train a molecular classifier for improving its generalization capability. On eight benchmark datasets we empirically demonstrate the effectiveness of MDP on the weakly supervised molecule classification tasks.
****
RadSimReal: Bridging the Gap Between Synthetic and Real Data in Radar Object Detection With Simulation
Oded Bialer, Yuval Haitman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15407-15416
Object detection in radar imagery with neural networks shows great potential for

improving autonomous driving. However obtaining annotated datasets from real radar images crucial for training these networks is challenging especially in scenarios with long-range detection and adverse weather and lighting conditions where radar performance excels. To address this challenge we present RadSimReal an innovative physical radar simulation capable of generating synthetic radar images with accompanying annotations for various radar types and environmental conditions all without the need for real data collection. Remarkably our findings demonstrate that training object detection models on RadSimReal data and subsequently evaluating them on real-world data produce performance levels comparable to models trained and tested on real data from the same dataset and even achieves better performance when testing across different real datasets. RadSimReal offers advantages over other physical radar simulations that it does not necessitate knowledge of the radar design details which are often not disclosed by radar suppliers and has faster run-time. This innovative tool has the potential to advance the development of computer vision algorithms for radar-based autonomous driving applications.
****
No More Ambiguity in 360deg Room Layout via Bi-Layout Estimation
Yu-Ju Tsai, Jin-Cheng Jhang, Jingjing Zheng, Wei Wang, Albert Y. C. Chen, Min Sun, Cheng-Hao Kuo, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28056-28065
Inherent ambiguity in layout annotations poses significant challenges to developing accurate 360deg room layout estimation models. To address this issue we propose a novel Bi-Layout model capable of predicting two distinct layout types. One stops at ambiguous regions while the other extends to encompass all visible areas. Our model employs two global context embeddings where each embedding is designed to capture specific contextual information for each layout type. With our novel feature guidance module the image feature retrieves relevant context from these embeddings generating layout-aware features for precise bi-layout predictions. A unique property of our Bi-Layout model is its ability to inherently detect ambiguous regions by comparing the two predictions. To circumvent the need for manual correction of ambiguous annotations during testing we also introduce a new metric for disambiguating ground truth layouts. Our method demonstrates superior performance on benchmark datasets notably outperforming leading approaches. Specifically on the MatterportLayout dataset it improves 3DIoU from 81.70% to 82.57% across the full test set and notably from 54.80% to 59.97% in subsets with significant ambiguity.
****
Residual Denoising Diffusion Models
Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, Liangqiong Qu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2773-2783
We propose residual denoising diffusion models (RDDM) a novel dual diffusion process that decouples the traditional single denoising diffusion process into residual diffusion and noise diffusion. This dual diffusion framework expands the denoising-based diffusion models initially uninterpretable for image restoration into a unified and interpretable model for both image generation and restoration by introducing residuals. Specifically our residual diffusion represents directional diffusion from the target image to the degraded input image and explicitly guides the reverse generation process for image restoration while noise diffusion represents random perturbations in the diffusion process. The residual prioritizes certainty while the noise emphasizes diversity enabling RDDM to effectively unify tasks with varying certainty or diversity requirements such as image generation and restoration. We demonstrate that our sampling process is consistent with that of DDPM and DDIM through coefficient transformation and propose a partially path-independent generation process to better understand the reverse process. Notably our RDDM enables a generic UNet trained with only an L1 loss and a batch size of 1 to compete with state-of-the-art image restoration methods. We provide code and pre-trained models to encourage further exploration application and development of our innovative framework (https://github.com/nachifur/RDDM).

****
Towards Accurate and Robust Architectures via Neural Architecture Search
Yuwei Ou, Yuqi Feng, Yanan Sun; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2024, pp. 5967-5976
To defend deep neural networks from adversarial attacks adversarial training has
 been drawing increasing attention for its effectiveness. However the accuracy a
nd robustness resulting from the adversarial training are limited by the archite
cture because adversarial training improves accuracy and robustness by adjusting
 the weight connection affiliated to the architecture. In this work we propose A
RNAS to search for accurate and robust architectures for adversarial training. F
irst we design an accurate and robust search space in which the placement of the
 cells and the proportional relationship of the filter numbers are carefully det
ermined. With the design the architectures can obtain both accuracy and robustne
ss by deploying accurate and robust structures to their sensitive positions resp
ectively. Then we propose a differentiable multi-objective search strategy perfo
rming gradient descent towards directions that are beneficial for both natural l
oss and adversarial loss thus the accuracy and robustness can be guaranteed at t
he same time. We conduct comprehensive experiments in terms of white-box attacks
 black-box attacks and transferability. Experimental results show that the searc
hed architecture has the strongest robustness with the competitive accuracy and
breaks the traditional idea that NAS-based architectures cannot transfer well to
 complex tasks in robustness scenarios. By analyzing outstanding architectures s
earched we also conclude that accurate and robust neural architectures tend to d
eploy different structures near the input and output which has great practical s
ignificance on both hand-crafting and automatically designing of accurate and ro
bust architectures.
****
Closely Interactive Human Reconstruction with Proxemics and Physics-Guided Adapt
ion
Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, Gim Hee Lee; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 1011-1021
Existing multi-person human reconstruction approaches mainly focus on recovering
 accurate poses or avoiding penetration but overlook the modeling of close inter
actions. In this work we tackle the task of reconstructing closely interactive h
umans from a monocular video. The main challenge of this task comes from insuffi
cient visual information caused by depth ambiguity and severe inter-person occlu
sion. In view of this we propose to leverage knowledge from proxemic behavior an
d physics to compensate the lack of visual information. This is based on the obs
ervation that human interaction has specific patterns following the social proxe
mics. Specifically we first design a latent representation based on Vector Quant
ised-Variational AutoEncoder (VQ-VAE) to model human interaction. A proxemics an
d physics guided diffusion model is then introduced to denoise the initial distr
ibution. We design the diffusion model as dual branch with each branch represent
ing one individual such that the interaction can be modeled via cross attention.
 With the learned priors of VQ-VAE and physical constraint as the additional inf
ormation our proposed approach is capable of estimating accurate poses that are
also proxemics and physics plausible. Experimental results on Hi4D 3DPW and CHI3
D demonstrate that our method outperforms existing approaches. The code is avail
able at https://github.com/boycehbz/HumanInteraction.
****
A Noisy Elephant in the Room: Is Your Out-of-Distribution Detector Robust to Lab
el Noise?
Galadrielle Humblot-Renaux, Sergio Escalera, Thomas B. Moeslund; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 22626-22636
The ability to detect unfamiliar or unexpected images is essential for safe depl
oyment of computer vision systems. In the context of classification the task of
detecting images outside of a model's training domain is known as out-of-distrib
ution (OOD) detection. While there has been a growing research interest in devel

oping post-hoc OOD detection methods there has been comparably little discussion around how these methods perform when the underlying classifier is not trained on a clean carefully curated dataset. In this work we take a closer look at 20 state-of-the-art OOD detection methods in the (more realistic) scenario where the labels used to train the underlying classifier are unreliable (e.g. crowd-sourced or web-scraped labels). Extensive experiments across different datasets noise types & levels architectures and checkpointing strategies provide insights into the effect of class label noise on OOD detection and show that poor separation between incorrectly classified ID samples vs. OOD samples is an overlooked yet important limitation of existing methods. Code: https://github.com/glhr/ood-labelnoise
****
VideoMAC: Video Masked Autoencoders Meet ConvNets
Gensheng Pei, Tao Chen, Xiruo Jiang, Huafeng Liu, Zeren Sun, Yazhou Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22733-22743
Recently the advancement of self-supervised learning techniques like masked autoencoders (MAE) has greatly influenced visual representation learning for images and videos. Nevertheless it is worth noting that the predominant approaches in existing masked image / video modeling rely excessively on resource-intensive vision transformers (ViTs) as the feature encoder. In this paper we propose a new approach termed as VideoMAC which combines video masked autoencoders with resource-friendly ConvNets. Specifically VideoMAC employs symmetric masking on randomly sampled pairs of video frames. To prevent the issue of mask pattern dissipation we utilize ConvNets which are implemented with sparse convolutional operators as encoders. Simultaneously we present a simple yet effective masked video modeling (MVM) approach a dual encoder architecture comprising an online encoder and an exponential moving average target encoder aimed to facilitate inter-frame reconstruction consistency in videos. Additionally we demonstrate that VideoMAC empowering classical (ResNet) / modern (ConvNeXt) convolutional encoders to harness the benefits of MVM outperforms ViT-based approaches on downstream tasks including video object segmentation (+5.2% / 6.4% \mathcal J &\mathcal F) body part propagation (+6.3% / 3.1% mIoU) and human pose tracking (+10.2% / 11.1% PCK@0.1).
****
Taming Stable Diffusion for Text to 360 Panorama Image Generation
Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6347-6357
Generative models e.g. Stable Diffusion have enabled the creation of photorealistic images from text prompts. Yet the generation of 360-degree panorama images from text remains a challenge particularly due to the dearth of paired text-panorama data and the domain gap between panorama and perspective images. In this paper we introduce a novel dual-branch diffusion model named PanFusion to generate a 360-degree image from a text prompt. We leverage the stable diffusion model as one branch to provide prior knowledge in natural image generation and register it to another panorama branch for holistic image generation. We propose a unique cross-attention mechanism with projection awareness to minimize distortion during the collaborative denoising process. Our experiments validate that PanFusion surpasses existing methods and thanks to its dual-branch structure can integrate additional constraints like room layout for customized panorama outputs.
****
3DSFLabelling: Boosting 3D Scene Flow Estimation by Pseudo Auto-labelling
Chaokang Jiang, Guangming Wang, Jiuming Liu, Hesheng Wang, Zhuang Ma, Zhenqiang Liu, Zhujin Liang, Yi Shan, Dalong Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15173-15183
Learning 3D scene flow from LiDAR point clouds presents significant difficulties including poor generalization from synthetic datasets to real scenes scarcity of real-world 3D labels and poor performance on real sparse LiDAR point clouds. We present a novel approach from the perspective of auto-labelling aiming to generate a large number of 3D scene flow pseudo labels for real-world LiDAR point cl

ouds. Specifically we employ the assumption of rigid body motion to simulate potential object-level rigid movements in autonomous driving scenarios. By updating different motion attributes for multiple anchor boxes the rigid motion decomposition is obtained for the whole scene. Furthermore we developed a novel 3D scene flow data augmentation method for global and local motion. By perfectly synthesizing target point clouds based on augmented motion parameters we easily obtain lots of 3D scene flow labels in point clouds highly consistent with real scenarios. On multiple real-world datasets including LiDAR KITTI nuScenes and Argoverse our method outperforms all previous supervised and unsupervised methods without requiring manual labelling. Impressively our method achieves a tenfold reduction in EPE3D metric on the LiDAR KITTI dataset reducing it from 0.190m to a mere 0.008m error.
****
Unsigned Orthogonal Distance Fields: An Accurate Neural Implicit Representation for Diverse 3D Shapes
Yujie Lu, Long Wan, Nayu Ding, Yulong Wang, Shuhan Shen, Shen Cai, Lin Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20551-20560
Neural implicit representation of geometric shapes has witnessed considerable advancements in recent years. However common distance field based implicit representations specifically signed distance field (SDF) for watertight shapes or unsigned distance field (UDF) for arbitrary shapes routinely suffer from degradation of reconstruction accuracy when converting to explicit surface points and meshes. In this paper we introduce a novel neural implicit representation based on unsigned orthogonal distance fields (UODFs). In UODFs the minimal unsigned distance from any spatial point to the shape surface is defined solely in one orthogonal direction contrasting with the multi-directional determination made by SDF and UDF. Consequently every point in the 3D UODFs can directly access its closest surface points along three orthogonal directions. This distinctive feature leverages the accurate reconstruction of surface points without interpolation errors. We verify the effectiveness of UODFs through a range of reconstruction examples extending from simple watertight or non-watertight shapes to complex shapes that include hollows internal or assembling structures.
****
Modular Blind Video Quality Assessment
Wen Wen, Mu Li, Yabin Zhang, Yiting Liao, Junlin Li, Li Zhang, Kede Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2763-2772
Blind video quality assessment (BVQA) plays a pivotal role in evaluating and improving the viewing experience of end-users across a wide range of video-based platforms and services. Contemporary deep learning-based models primarily analyze video content in its aggressively subsampled format while being blind to the impact of the actual spatial resolution and frame rate on video quality. In this paper we propose a modular BVQA model and a method of training it to improve its modularity. Our model comprises a base quality predictor a spatial rectifier and a temporal rectifier responding to the visual content and distortion spatial resolution and frame rate changes on video quality respectively. During training spatial and temporal rectifiers are dropped out with some probabilities to render the base quality predictor a standalone BVQA model which should work better with the rectifiers. Extensive experiments on both professionally-generated content and user-generated content video databases show that our quality model achieves superior or comparable performance to current methods. Additionally the modularity of our model offers an opportunity to analyze existing video quality databases in terms of their spatial and temporal complexity.
****
Question Aware Vision Transformer for Multimodal Reasoning
Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, Ron Litman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13861-13871
Vision-Language (VL) models have gained significant research focus enabling rema

rkable advances in multimodal reasoning. These architectures typically comprise a vision encoder a Large Language Model (LLM) and a projection module that aligns visual features with the LLM's representation space. Despite their success a critical limitation persists: the vision encoding process remains decoupled from user queries often in the form of image-related questions. Consequently the resulting visual features may not be optimally attuned to the query-specific elements of the image. To address this we introduce QA-ViT a Question Aware Vision Transformer approach for multimodal reasoning which embeds question awareness directly within the vision encoder. This integration results in dynamic visual features focusing on relevant image aspects to the posed question. QA-ViT is model-agnostic and can be incorporated efficiently into any VL architecture. Extensive experiments demonstrate the effectiveness of applying our method to various multimodal architectures leading to consistent improvement across diverse tasks and showcasing its potential for enhancing visual and scene-text understanding.
****
OST: Refining Text Knowledge with Optimal Spatio-Temporal Descriptor for General Video Recognition
Tongjia Chen, Hongshan Yu, Zhengeng Yang, Zechuan Li, Wei Sun, Chen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18888-18898
Due to the resource-intensive nature of training vision-language models on expansive video data a majority of studies have centered on adapting pre-trained image-language models to the video domain. Dominant pipelines propose to tackle the visual discrepancies with additional temporal learners while overlooking the substantial discrepancy for web-scaled descriptive narratives and concise action category names leading to less distinct semantic space and potential performance limitations. In this work we prioritize the refinement of text knowledge to facilitate generalizable video recognition. To address the limitations of the less distinct semantic space of category names we prompt a large language model (LLM) to augment action class names into Spatio-Temporal Descriptors thus bridging the textual discrepancy and serving as a knowledge base for general recognition. Moreover to assign the best descriptors with different video instances we propose Optimal Descriptor Solver forming the video recognition problem as solving the optimal matching flow across frame-level representations and descriptors. Comprehensive evaluations in zero-shot few-shot and fully supervised video recognition highlight the effectiveness of our approach. Our best model achieves a state-of-the-art zero-shot accuracy of 75.1% on Kinetics-600.
****
Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation
Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, Manolis Savva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16384-16393
We contribute the Habitat Synthetic Scene Dataset a dataset of 211 high-quality 3D scenes and use it to test navigation agent generalization to realistic 3D environments. Our dataset represents real interiors and contains a diverse set of 18656 models of real-world objects. We investigate the impact of synthetic 3D scene dataset scale and realism on the task of training embodied agents to find and navigate to objects (ObjectGoal navigation). By comparing to synthetic 3D scene datasets from prior work we find that scale helps in generalization but the benefits quickly saturate making visual fidelity and correlation to real-world scenes more important. Our experiments show that agents trained on our smaller-scale dataset can match or outperform agents trained on much larger datasets. Surprisingly we observe that agents trained on just 122 scenes from our dataset outperform agents trained on 10000 scenes from the ProcTHOR-10K dataset in terms of zero-shot generalization in real-world scanned environments.
****
OA-CNNs: Omni-Adaptive Sparse CNNs for 3D Semantic Segmentation
Bohao Peng, Xiaoyang Wu, Li Jiang, Yukang Chen, Hengshuang Zhao, Zhuotao Tian, J

iaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21305-21315
The booming of 3D recognition in the 2020s began with the introduction of point cloud transformers. They quickly overwhelmed sparse CNNs and became state-of-the-art models especially in 3D semantic segmentation. However sparse CNNs are still valuable networks due to their efficiency treasure and ease of application. In this work we reexamine the design distinctions and test the limits of what a sparse CNN can achieve. We discover that the key credit to the performance difference is adaptivity. Specifically we propose two key components i.e. adaptive receptive fields (spatially) and adaptive relation to bridge the gap. This exploration led to the creation of Omni-Adaptive 3D CNNs (OA-CNNs) a family of networks that integrates a lightweight module to greatly enhance the adaptivity of sparse CNNs at minimal computational cost. Without any self-attention modules OA-CNNs favorably surpass point transformers in terms of accuracy in both indoor and outdoor scenes with much less latency and memory cost. Notably it achieves 76.1% 78.9% and 70.6% mIoU on ScanNet v2 nuScenes and SemanticKITTI validation benchmarks respectively while maintaining at most 5x better speed than transformer counterparts. This revelation highlights the potential of pure sparse CNNs to outperform transformer-related networks. Our code is built upon Pointcept which is available at https://github.com/Pointcept/Pointcept.
****
RELI11D: A Comprehensive Multimodal Human Motion Dataset and Method
Ming Yan, Yan Zhang, Shuqiang Cai, Shuqi Fan, Xincheng Lin, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2250-2262
Comprehensive capturing of human motions requires both accurate captures of complex poses and precise localization of the human within scenes. Most of the HPE datasets and methods primarily rely on RGB LiDAR or IMU data. However solely using these modalities or a combination of them may not be adequate for HPE particularly for complex and fast movements. For holistic human motion understanding we present RELI11D a high-quality multimodal human motion dataset involves LiDAR IMU system RGB camera and Event camera. It records the motions of 10 actors performing 5 sports in 7 scenes including 3.32 hours of synchronized LiDAR point clouds IMU measurement data RGB videos and Event steams. Through extensive experiments we demonstrate that the RELI11D presents considerable challenges and opportunities as it contains many rapid and complex motions that require precise location. To address the challenge of integrating different modalities we propose LEIR a multimodal baseline that effectively utilizes LiDAR Point Cloud Event stream and RGB through our cross-attention fusion strategy. We show that LEIR exhibits promising results for rapid motions and daily motions and that utilizing the characteristics of multiple modalities can indeed improve HPE performance. Both the dataset and source code will be released publicly to the research community fostering collaboration and enabling further exploration in this field.
****
Generative Image Dynamics
Zhengqi Li, Richard Tucker, Noah Snavely, Aleksander Holynski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24142-24153
We present an approach to modeling an image-space prior on scene motion. Our prior is learned from a collection of motion trajectories extracted from real video sequences depicting natural oscillatory dynamics of objects such as treesflowers candles and clothes swaying in the wind. We model dense long-term motion in the Fourier domain as spectral volumes which we find are well-suited to prediction with diffusion models. Given a single image our trained model uses a frequency-coordinated diffusion sampling process to predict a spectral volume which can be converted into a motion texture that spans an entire video. Along with an image-based rendering module the predicted motion representation can be used for a number of downstream applications such as turning still images into seamlessly looping videos or allowing users to interact with objects in real images producing realistic simulated dynamics (by interpreting the spectral volumes as image-spac

e modal bases). See our project page for more results: generative-dynamics.githu
b.io
****
One-Class Face Anti-spoofing via Spoof Cue Map-Guided Feature Learning
Pei-Kai Huang, Cheng-Hsuan Chiang, Tzu-Hsien Chen, Jun-Xiong Chong, Tyng-Luh Liu
, Chiou-Ting Hsu; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2024, pp. 277-286
Many face anti-spoofing (FAS) methods have focused on learning discriminative fe
atures from both live and spoof training data to strengthen the security of face
 recognition systems. However since not every possible attack type is available
in the training stage these FAS methods usually fail to detect unseen attacks in
 the inference stage. In comparison one-class FAS where the training data are fr
om only live faces aims to detect whether a test face image belongs to the live
class or not. In this paper we propose a novel One-Class Spoof Cue Map estimatio
n Network (OC-SCMNet) to address the one-class FAS detection problem. Our first
goal is to learn to extract latent spoof features from live images so that their
 estimated Spoof Cue Maps (SCMs) should have zero responses. To avoid trapping t
o a trivial solution we devise a novel SCM-guided feature learning by combining
many SCMs as pseudo ground-truths to guide a conditional generator to generate l
atent spoof features for spoof data. Our second goal is to approximately simulat
e the potential out-of-distribution spoof attacks. To this end we propose using
a memory bank to dynamically preserve a set of sufficiently "independent" latent
 spoof features to encourage the generator to probe the latent spoof feature spa
ce. Extensive experiments conducted on eight FAS benchmark datasets demonstrate
that the proposed OC-SCMNet not only outperforms previous one-class FAS methods
but also achieves comparable performances to state-of-the-art two-class FAS meth
od. The codes are available at https://github.com/Pei-KaiHuang/CVPR24_OC_SCMNet.
****
On the Test-Time Zero-Shot Generalization of Vision-Language Models: Do We Reall
y Need Prompt Learning?
Maxime Zanella, Ismail Ben Ayed; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2024, pp. 23783-23793
The development of large vision-language models notably CLIP has catalyzed resea
rch into effective adaptation techniques with a particular focus on soft prompt
tuning. Conjointly test-time augmentation which utilizes multiple augmented view
s of a single image to enhance zero-shot generalization is emerging as a signifi
cant area of interest. This has predominantly directed research efforts towards
test-time prompt tuning. In contrast we introduce a robust MeanShift for Test-ti
me Augmentation (MTA) which surpasses prompt-based methods without requiring thi
s intensive training procedure. This positions MTA as an ideal solution for both
 standalone and API-based applications. Additionally our method does not rely on
 ad hoc rules (e.g. confidence threshold) used in some previous test-time augmen
tation techniques to filter the augmented views. Instead MTA incorporates a qual
ity assessment variable for each view directly into its optimization process ter
med as the inlierness score. This score is jointly optimized with a density mode
 seeking process leading to an efficient training- and hyperparameter-free appro
ach. We extensively benchmark our method on 15 datasets and demonstrate MTA's su
periority and computational efficiency. Deployed easily as plug-and-play module
on top of zero-shot models and state-of-the-art few-shot methods MTA shows syste
matic and consistent improvements.
****
InteractDiffusion: Interaction Control in Text-to-Image Diffusion Models
Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, Weipeng Hu; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2024, pp. 6180-6189
Large-scale text-to-image (T2I) diffusion models have showcased incredible capab
ilities in generating coherent images based on textual descriptions enabling vas
t applications in content generation. While recent advancements have introduced
control over factors such as object localization posture and image contours a cr
ucial gap remains in our ability to control the interactions between objects in

the generated content. Well-controlling interactions in generated images could y
ield meaningful applications such as creating realistic scenes with interacting
characters. In this work we study the problems of conditioning T2I diffusion mod
els with Human-Object Interaction (HOI) information consisting of a triplet labe
l (person action object) and corresponding bounding boxes. We propose a pluggabl
e interaction control model called InteractDiffusion that extends existing pre-t
rained T2I diffusion models to enable them being better conditioned on interacti
ons. Specifically we tokenize the HOI information and learn their relationships
via interaction embeddings. A conditioning self-attention layer is trained to ma
p HOI tokens to visual tokens thereby conditioning the visual tokens better in e
xisting T2I diffusion models. Our model attains the ability to control the inter
action and location on existing T2I diffusion models which outperforms existing
baselines by a large margin in HOI detection score as well as fidelity in FID an
d KID. Project page: https://jiuntian.github.io/interactdiffusion.
****
NViST: In the Wild New View Synthesis from a Single Image with Transformers
Wonbong Jang, Lourdes Agapito; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2024, pp. 10181-10193
We propose NViST a transformer-based model for efficient and generalizable novel
-view synthesis from a single image for real-world scenes. In contrast to many m
ethods that are trained on synthetic data object-centred scenarios or in a categ
ory-specific manner NViST is trained on MVImgNet a large-scale dataset of casual
ly-captured real-world videos of hundreds of object categories with diverse back
grounds. NViST transforms image inputs directly into a radiance field conditione
d on camera parameters via adaptive layer normalisation. In practice NViST explo
its fine-tuned masked autoencoder (MAE) features and translates them to 3D outpu
t tokens via cross-attention while addressing occlusions with self-attention. To
 move away from object-centred datasets and enable full scene synthesis NViST ad
opts a 6-DOF camera pose model and only requires relative pose dropping the need
 for canonicalization of the training data which removes a substantial barrier t
o it being used on casually captured datasets. We show results on unseen objects
 and categories from MVImgNet and even generalization to casual phone captures.
We conduct qualitative and quantitative evaluations on MVImgNet and ShapeNet to
show that our model represents a step forward towards enabling true in-the-wild
generalizable novel-view synthesis from a single image. Project webpage: https:/
/wbjang.github.io/nvist_webpage.
****
Beyond Text: Frozen Large Language Models in Visual Signal Comprehension
Lei Zhu, Fangyun Wei, Yanye Lu; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2024, pp. 27047-27057
In this work we investigate the potential of a large language model (LLM) to dir
ectly comprehend visual signals without the necessity of fine-tuning on multi-mo
dal datasets. The foundational concept of our method views an image as a linguis
tic entity and translates it to a set of discrete words derived from the LLM's v
ocabulary. To achieve this we present the Vision-to-Language Tokenizer abbreviat
ed as V2T Tokenizer which transforms an image into a "foreign language" with the
 combined aid of an encoder-decoder the LLM vocabulary and a CLIP model. With th
is innovative image encoding the LLM gains the ability not only for visual compr
ehension but also for image denoising and restoration in an auto-regressive fash
ion--crucially without any fine-tuning. We undertake rigorous experiments to val
idate our method encompassing understanding tasks like image recognition image c
aptioning and visual question answering as well as image denoising tasks like in
painting outpainting deblurring and shift restoration. Code and models are avail
able at https://github.com/zh460045050/V2L-Tokenizer.
****
Rotated Multi-Scale Interaction Network for Referring Remote Sensing Image Segme
ntation
Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, Rongr
ong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2024, pp. 26658-26668

Referring Remote Sensing Image Segmentation (RRSIS) is a new challenge that comb
ines computer vision and natural language processing. Traditional Referring Imag
e Segmentation (RIS) approaches have been impeded by the complex spatial scales
and orientations found in aerial imagery leading to suboptimal segmentation resu
lts. To address these challenges we introduce the Rotated Multi-Scale Interactio
n Network (RMSIN) an innovative approach designed for the unique demands of RRSI
S. RMSIN incorporates an Intra-scale Interaction Module (IIM) to effectively add
ress the fine-grained detail required at multiple scales and a Cross-scale Inter
action Module (CIM) for integrating these details coherently across the network.
 Furthermore RMSIN employs an Adaptive Rotated Convolution (ARC) to account for
the diverse orientations of objects a novel contribution that significantly enha
nces segmentation accuracy. To assess the efficacy of RMSIN we have curated an e
xpansive dataset comprising 17402 image-caption-mask triplets which is unparalle
led in terms of scale and variety. This dataset not only presents the model with
 a wide range of spatial and rotational scenarios but also establishes a stringe
nt benchmark for the RRSIS task ensuring a rigorous evaluation of performance. E
xperimental evaluations demonstrate the exceptional performance of RMSIN surpass
ing existing state-of-the-art models by a significant margin. Datasets and code
are available at https://github.com/Lsan2401/RMSIN.
****
GLACE: Global Local Accelerated Coordinate Encoding
Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, Marc Pollefeys
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2024, pp. 21562-21571
Scene coordinate regression (SCR) methods are a family of visual localization me
thods that directly regress 2D-3D matches for camera pose estimation. They are e
ffective in small-scale scenes but face significant challenges in large-scale sc
enes that are further amplified in the absence of ground truth 3D point clouds f
or supervision. Here the model can only rely on reprojection constraints and nee
ds to implicitly triangulate the points. The challenges stem from a fundamental
dilemma: The network has to be invariant to observations of the same landmark at
 different viewpoints and lighting conditions etc. but at the same time discrimi
nate unrelated but similar observations. The latter becomes more relevant and se
vere in larger scenes. In this work we tackle this problem by introducing the co
ncept of co-visibility to the network. We propose GLACE which integrates pre-tra
ined global and local encodings and enables SCR to scale to large scenes with on
ly a single small-sized network. Specifically we propose a novel feature diffusi
on technique that implicitly groups the reprojection constraints with co-visibil
ity and avoids overfitting to trivial solutions. Additionally our position decod
er parameterizes the output positions for large-scale scenes more effectively. W
ithout using 3D models or depth maps for supervision our method achieves state-o
f-the-art results on large-scale scenes with a low-map-size model. On Cambridge
landmarks with a single model we achieve a 17% lower median position error than
Poker the ensemble variant of the state-of-the-art SCR method ACE. Code is avail
able at: https://github.com/cvg/glace.
****
Emergent Open-Vocabulary Semantic Segmentation from Off-the-shelf Vision-Languag
e Models
Jiayun Luo, Siddhesh Khandelwal, Leonid Sigal, Boyang Li; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 40
29-4040
From image-text pairs large-scale vision-language models (VLMs) learn to implici
tly associate image regions with words which prove effective for tasks like visu
al question answering. However leveraging the learned association for open-vocab
ulary semantic segmentation remains a challenge. In this paper we propose a simp
le yet extremely effective training-free technique Plug-and-Play Open-Vocabulary
 Semantic Segmentation (PnP-OVSS) for this task. PnP-OVSS leverages a VLM with d
irect text-to-image cross-attention and an image-text matching loss. To balance
between over-segmentation and under-segmentation we introduce Salience Dropout;
by iteratively dropping patches that the model is most attentive to we are able

to better resolve the entire extent of the segmentation mask. PnP-OVSS does not require any neural network training and performs hyperparameter tuning without the need for any segmentation annotations even for a validation set. PnP-OVSS demonstrates substantial improvements over comparable baselines (+29.4% mIoU on Pascal VOC +13.2% mIoU on Pascal Context +14.0% mIoU on MS COCO +2.4% mIoU on COCO Stuff) and even outperforms most baselines that conduct additional network training on top of pretrained VLMs. Our codebase is at https://github.com/letitiabanana/PnP-OVSS.
****
Localization Is All You Evaluate: Data Leakage in Online Mapping Datasets and How to Fix It
Adam Lilja, Junsheng Fu, Erik Stenborg, Lars Hammarstrand; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22150-22159
The task of online mapping is to predict a local map using current sensor observations e.g. from lidar and camera without relying on a pre-built map. State-of-the-art methods are based on supervised learning and are trained predominantly using two datasets: nuScenes and Argoverse 2. However these datasets revisit the same geographic locations across training validation and test sets. Specifically over 80% of nuScenes and 40% of Argoverse 2 validation and test samples are less than 5 m from a training sample. At test time the methods are thus evaluated more on how well they localize within a memorized implicit map built from the training data than on extrapolating to unseen locations. Naturally this data leakage causes inflated performance numbers and we propose geographically disjoint data splits to reveal the true performance in unseen environments. Experimental results show that methods perform considerably worse some dropping more than 45 mAP when trained and evaluated on proper data splits. Additionally a reassessment of prior design choices reveals diverging conclusions from those based on the original split. Notably the impact of lifting methods and the support from auxiliary tasks (e.g. depth supervision) on performance appears less substantial or follows a different trajectory than previously perceived.
****
Alchemist: Parametric Control of Material Properties with Diffusion Models
Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, Bill Freeman, Mark Matthews; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24130-24141
We propose a method to control material attributes of objects like roughness metallic albedo and transparency in real images. Our method capitalizes on the generative prior of text-to-image models known for photorealism employing a scalar value and instructions to alter low-level material properties. Addressing the lack of datasets with controlled material attributes we generated an object-centric synthetic dataset with physically-based materials. Fine-tuning a modified pre-trained text-to-image model on this synthetic dataset enables us to edit material properties in real-world images while preserving all other attributes. We show the potential application of our model to material edited NeRFs.
****
Step Differences in Instructional Video
Tushar Nagarajan, Lorenzo Torresani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18740-18750
Comparing a user video to a reference how-to video is a key requirement for AR/VR technology delivering personalized assistance tailored to the user's progress. However current approaches for language-based assistance can only answer questions about a single video. We propose an approach that first automatically generates large amounts of visual instruction tuning data involving pairs of videos from HowTo100M by leveraging existing step annotations and accompanying narrations and then trains a video-conditioned language model to jointly reason across multiple raw videos. Our model achieves state-of-the-art performance at identifying differences between video pairs and ranking videos based on the severity of these differences and shows promising ability to perform general reasoning over multiple videos.

****
Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data
Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, Hengshuang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2024, pp. 10371-10381
This work presents Depth Anything a highly practical solution for robust monocul
ar depth estimation. Without pursuing novel technical modules we aim to build a
simple yet powerful foundation model dealing with any images under any circumsta
nces. To this end we scale up the dataset by designing a data engine to collect
and automatically annotate large-scale unlabeled data (62M) which significantly
 enlarges the data coverage and thus is able to reduce the generalization error.
 We investigate two simple yet effective strategies that make data scaling-up pr
omising. First a more challenging optimization target is created by leveraging d
ata augmentation tools. It compels the model to actively seek extra visual knowl
edge and acquire robust representations. Second an auxiliary supervision is deve
loped to enforce the model to inherit rich semantic priors from pre-trained enco
ders. We evaluate its zero-shot capabilities extensively including six public da
tasets and randomly captured photos. It demonstrates impressive generalization a
bility. Further through fine-tuning it with metric depth information from NYUv2
and KITTI new SOTAs are set. Our better depth model also results in a better dep
th-conditioned ControlNet. Our models are released at https://github.com/LiheYou
ng/Depth-Anything.
****
SelfPose3d: Self-Supervised Multi-Person Multi-View 3d Pose Estimation
Vinkle Srivastav, Keqi Chen, Nicolas Padoy; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2502-2512
We present a new self-supervised approach SelfPose3d for estimating 3d poses of
multiple persons from multiple camera views. Unlike current state-of-the-art ful
ly-supervised methods our approach does not require any 2d or 3d ground-truth po
ses and uses only the multi-view input images from a calibrated camera setup and
 2d pseudo poses generated from an off-the-shelf 2d human pose estimator. We pro
pose two self-supervised learning objectives: self-supervised person localizatio
n in 3d space and self-supervised 3d pose estimation. We achieve self-supervised
 3d person localization by training the model on synthetically generated 3d poin
ts serving as 3d person root positions and on the projected root-heatmaps in all
 the views. We then model the 3d poses of all the localized persons with a bottl
eneck representation map them onto all views obtaining 2d joints and render them
 using 2d Gaussian heatmaps in an end-to-end differentiable manner. Afterwards w
e use the corresponding 2d joints and heatmaps from the pseudo 2d poses for lear
ning. To alleviate the intrinsic inaccuracy of the pseudo labels we propose an a
daptive supervision attention mechanism to guide the self-supervision. Our exper
iments and analysis on three public benchmark datasets including Panoptic Shelf
and Campus show the effectiveness of our approach which is comparable to fully-s
upervised methods. Code is available at https://github.com/CAMMA-public/SelfPose
3D.
****
MoDE: CLIP Data Experts via Clustering
Jiawei Ma, Po-Yao Huang, Saining Xie, Shang-Wen Li, Luke Zettlemoyer, Shih-Fu Ch
ang, Wen-Tau Yih, Hu Xu; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 26354-26363
The success of contrastive language-image pretraining (CLIP) relies on the super
vision from the pairing between images and captions which tends to be noisy in w
eb-crawled data. We present Mixture of Data Experts (MoDE) and learn a system of
 CLIP data experts via clustering. Each data expert is trained on one data clust
er being less sensitive to false negative noises in other clusters. At inference
 time we ensemble their outputs by applying weights determined through the corre
lation between task metadata and cluster conditions. To estimate the correlation
 precisely the samples in one cluster should be semantically similar but the num
ber of data experts should still be reasonable for training and inference. As su
ch we consider the ontology in human language and propose to use fine-grained cl

uster centers to represent each data expert at a coarse-grained level. Experimen
tal studies show that four CLIP data experts on ViT-B/16 outperform the ViT-L/14
 by OpenAI CLIP and OpenCLIP on zero-shot image classification but with less (<3
5%) training cost. Meanwhile MoDE can train all data expert asynchronously and c
an flexibly include new data experts. The code is available here.
****
Joint2Human: High-Quality 3D Human Generation via Compact Spherical Embedding of
 3D Joints
Muxin Zhang, Qiao Feng, Zhuo Su, Chao Wen, Zhou Xue, Kun Li; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 1429-1438
3D human generation is increasingly significant in various applications. However
 the direct use of 2D generative methods in 3D generation often results in losin
g local details while methods that reconstruct geometry from generated images st
ruggle with global view consistency. In this work we introduce Joint2Human a nov
el method that leverages 2D diffusion models to generate detailed 3D human geome
try directly ensuring both global structure and local details. To achieve this w
e employ the Fourier occupancy field (FOF) representation enabling the direct ge
neration of 3D shapes as preliminary results with 2D generative models. With the
 proposed high-frequency enhancer and the multi-view recarving strategy our meth
od can seamlessly integrate the details from different views into a uniform glob
al shape. To better utilize the 3D human prior and enhance control over the gene
rated geometry we introduce a compact spherical embedding of 3D joints. This all
ows for an effective guidance of pose during the generation process. Additionall
y our method can generate 3D humans guided by textual inputs. Our experimental r
esults demonstrate the capability of our method to ensure global structure local
 details high resolution and low computational cost simultaneously. More results
 and the code can be found on our project page at http://cic.tju.edu.cn/faculty/
likun/projects/Joint2Human.
****
Prompt-Free Diffusion: Taking "Text" out of Text-to-Image Diffusion Models
Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, Humphrey Shi; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2024, pp. 8682-8692
Text-to-image (T2I) research has grown explosively in the past year owing to the
 large-scale pre-trained diffusion models and many emerging personalization and
editing approaches. Yet one pain point persists: the text prompt engineering and
 searching high-quality text prompts for customized results is more art than sci
ence. Moreover as commonly argued: "an image is worth a thousand words" - the at
tempt to describe a desired image with texts often ends up being ambiguous and c
annot comprehensively cover delicate visual details hence necessitating more add
itional controls from the visual domain. In this paper we take a bold step forwa
rd: taking "Text" out of a pretrained T2I diffusion model to reduce the burdenso
me prompt engineering efforts for users. Our proposed framework Prompt-Free Diff
usion relies on only visual inputs to generate new images: it takes a reference
image as "context" an optional image structural conditioning and an initial nois
e with absolutely no text prompt. The core architecture behind the scene is Sema
ntic Context Encoder (SeeCoder) substituting the commonly used CLIP-based or LLM
-based text encoder. The reusability of SeeCoder also makes it a convenient drop
-in component: one can also pre-train a SeeCoder in one T2I model and reuse it f
or another. Through extensive experiments Prompt-Free Diffusion is experimentall
y found to (i) outperform prior exemplar-based image synthesis approaches; (ii)
perform on par with state-of-the-art T2I models using prompts following the best
 practice; and (iii) be naturally extensible to other downstream applications su
ch as anime figure generation and virtual try-on with promising quality. Our cod
e and models will be open-sourced.
****
MPOD123: One Image to 3D Content Generation Using Mask-enhanced Progressive Outl
ine-to-Detail Optimization
Jimin Xu, Tianbao Wang, Tao Jin, Shengyu Zhang, Dongjie Fu, Zhe Wang, Jiangjing

Lyu, Chengfei Lv, Chaoyue Niu, Zhou Yu, Zhou Zhao, Fei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10682-10692
Recent advancements in single image driven 3D content generation have been propelled by leveraging prior knowledge from pretrained 2D diffusion models. However the 3D content generated by existing methods often exhibits distorted outline shapes and inadequate details. To solve this problem we propose a novel framework called Mask-enhanced Progressive Outline-to-Detail optimization (aka. MPOD123) which consists of two stages. Specifically in the first stage MPOD123 utilizes the pretrained view-conditioned diffusion model to guide the outline shape optimization of the 3D content. Given certain viewpoint we estimate outline shape priors in the form of 2D mask from the 3D content by leveraging opacity calculation. In the second stage MPOD123 incorporates Detail Appearance Inpainting (DAI) to guide the refinement on local geometry and texture with the shape priors. The essence of DAI lies in the Mask Rectified Cross-Attention (MRCA) which can be conveniently plugged in the stable diffusion model. The MRCA module utilizes the mask to rectify the attention map from each cross-attention layer. Accompanied with this new module DAI is capable of guiding the detail refinement of the 3D content while better preserves the outline shape. To assess the applicability in practical scenarios we contribute a new dataset modeled on real-world e-commerce environments. Extensive quantitative and qualitative experiments on this dataset and open benchmarks demonstrate the effectiveness of MPOD123 over the state-of-the-arts.
****
Multi-agent Long-term 3D Human Pose Forecasting via Interaction-aware Trajectory Conditioning
Jaewoo Jeong, Daehee Park, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1617-1628
Human pose forecasting garners attention for its diverse applications. However challenges in modeling the multi-modal nature of human motion and intricate interactions among agents persist particularly with longer timescales and more agents. In this paper we propose an interaction-aware trajectory-conditioned long-term multi-agent human pose forecasting model utilizing a coarse-to-fine prediction approach: multi-modal global trajectories are initially forecasted followed by respective local pose forecasts conditioned on each mode. In doing so our Trajectory2Pose model introduces a graph-based agent-wise interaction module for a reciprocal forecast of local motion-conditioned global trajectory and trajectory-conditioned local pose. Our model effectively handles the multi-modality of human motion and the complexity of long-term multi-agent interactions improving performance in complex environments. Furthermore we address the lack of long-term (6s+) multi-agent (5+) datasets by constructing a new dataset from real-world images and 2D annotations enabling a comprehensive evaluation of our proposed model. State-of-the-art prediction performance on both complex and simpler datasets confirms the generalized effectiveness of our method. The code is available at https://github.com/Jaewoo97/T2P.
****
UnionFormer: Unified-Learning Transformer with Multi-View Representation for Image Manipulation Detection and Localization
Shuaibo Li, Wei Ma, Jianwei Guo, Shibiao Xu, Benchong Li, Xiaopeng Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12523-12533
We present UnionFormer a novel framework that integrates tampering clues across three views by unified learning for image manipulation detection and localization. Specifically we construct a BSFI-Net to extract tampering features from RGB and noise views achieving enhanced responsiveness to boundary artifacts while modulating spatial consistency at different scales. Additionally to explore the inconsistency between objects as a new view of clues we combine object consistency modeling with tampering detection and localization into a three-task unified learning process allowing them to promote and improve mutually. Therefore we acquire a unified manipulation discriminative representation under multi-scale supervi

sion that consolidates information from three views. This integration facilitate
s highly effective concurrent detection and localization of tampering. We perfor
m extensive experiments on diverse datasets and the results show that the propos
ed approach outperforms state-of-the-art methods in tampering detection and loca
lization.
****
Situational Awareness Matters in 3D Vision Language Reasoning
Yunze Man, Liang-Yan Gui, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13678-13688
Being able to carry out complicated vision language reasoning tasks in 3D space
represents a significant milestone in developing household robots and human-cent
ered embodied AI. In this work we demonstrate that a critical and distinct chall
enge in 3D vision language reasoning is the situational awareness which incorpor
ates two key components: (1) The autonomous agent grounds its self-location base
d on a language prompt. (2) The agent answers open-ended questions from the pers
pective of its calculated position. To address this challenge we introduce SIG3D
 an end-to-end Situation-Grounded model for 3D vision language reasoning. We tok
enize the 3D scene into sparse voxel representation and propose a language-groun
ded situation estimator followed by a situated question answering module. Experi
ments on the SQA3D and ScanQA datasets show that SIG3D outperforms state-of-the-
art models in situational estimation and question answering by a large margin (e
.g. an enhancement of over 30% on situation accuracy). Subsequent analysis corro
borates our architectural design choices explores the distinct functions of visu
al and textual tokens and highlights the importance of situational awareness in
the domain of 3D question-answering. Project page is available at https://yunzem
an.github.io/situation3d.
****
RCBEVDet: Radar-camera Fusion in Bird's Eye View for 3D Object Detection
Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang
 Dong, Nan Dong, Le Zhang, Ce Zhu; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2024, pp. 14928-14937
Three-dimensional object detection is one of the key tasks in autonomous driving
. To reduce costs in practice low-cost multi-view cameras for 3D object detectio
n are proposed to replace the expansive LiDAR sensors. However relying solely on
 cameras is difficult to achieve highly accurate and robust 3D object detection.
 An effective solution to this issue is combining multi-view cameras with the ec
onomical millimeter-wave radar sensor to achieve more reliable multi-modal 3D ob
ject detection. In this paper we introduce RCBEVDet a radar-camera fusion 3D obj
ect detection method in the bird's eye view (BEV). Specifically we first design
RadarBEVNet for radar BEV feature extraction. RadarBEVNet consists of a dual-str
eam radar backbone and a Radar Cross-Section (RCS) aware BEV encoder. In the dua
l-stream radar backbone a point-based encoder and a transformer-based encoder ar
e proposed to extract radar features with an injection and extraction module to
facilitate communication between the two encoders. The RCS-aware BEV encoder tak
es RCS as the object size prior to scattering the point feature in BEV. Besides
we present the Cross-Attention Multi-layer Fusion module to automatically align
the multi-modal BEV feature from radar and camera with the deformable attention
mechanism and then fuse the feature with channel and spatial fusion layers. Expe
rimental results show that RCBEVDet achieves new state-of-the-art radar-camera f
usion results on nuScenes and view-of-delft (VoD) 3D object detection benchmarks
. Furthermore RCBEVDet achieves better 3D detection results than all real-time c
amera-only and radar-camera 3D object detectors with a faster inference speed at
 21 28 FPS. The source code will be released at https://github.com/VDIGPKU/RCBEV
Det.
****
CLOAF: CoLlisiOn-Aware Human Flow
Andrey Davydov, Martin Engilberge, Mathieu Salzmann, Pascal Fua; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 1176-1185
Even the best current algorithms for estimating body 3D shape and pose yield res

ults that include body self-intersections. In this paper we present CLOAF which exploits the diffeomorphic nature of Ordinary Differential Equations to eliminat e such self-intersections while still imposing body shape constraints. We show t hat unlike earlier approaches to addressing this issue ours completely eliminate s the self-intersections without compromising the accuracy of the reconstruction s. Being differentiable CLOAF can be used to fine-tune pose and shape estimation baselines to improve their overall performance and eliminate self-intersections in their predictions. Furthermore we demonstrate how our CLOAF strategy can be applied to practically any motion field induced by the user. CLOAF also makes it possible to edit motion to interact with the environment without worrying about potential collision or loss of body-shape prior.
****
Hybrid Functional Maps for Crease-Aware Non-Isometric Shape Matching
Lennart Bastian, Yizheng Xie, Nassir Navab, Zorah Lähner; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 33 13-3323
Non-isometric shape correspondence remains a fundamental challenge in computer v ision. Traditional methods using Laplace-Beltrami operator (LBO) eigenmodes face limitations in characterizing high-frequency extrinsic shape changes like bendi ng and creases. We propose a novel approach of combining the non-orthogonal extr insic basis of eigenfunctions of the elastic thin-shell hessian with the intrins ic ones of the LBO creating a hybrid spectral space in which we construct functi onal maps. To this end we present a theoretical framework to effectively integra te non-orthogonal basis functions into descriptor- and learning-based functional map methods. Our approach can be incorporated easily into existing functional m ap pipelines across varying applications and is able to handle complex deformati ons beyond isometries. We show extensive evaluations across various supervised a nd unsupervised settings and demonstrate significant improvements. Notably our a pproach achieves up to 15% better mean geodesic error for non-isometric correspo ndence settings and up to 45% improvement in scenarios with topological noise.
****
Density-Guided Semi-Supervised 3D Semantic Segmentation with Dual-Space Hardness Sampling
Jianan Li, Qiulei Dong; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2024, pp. 3260-3269
Densely annotating the large-scale point clouds is laborious. To alleviate the a nnotation burden contrastive learning has attracted increasing attention for tac kling semi-supervised 3D semantic segmentation. However existing point-to-point contrastive learning techniques in literature are generally sensitive to outlier s resulting in insufficient modeling of the point-wise representations. To addre ss this problem we propose a method named DDSemi for semi-supervised 3D semantic segmentation where a density-guided contrastive learning technique is explored. This technique calculates the contrastive loss in a point-to-anchor manner by e stimating an anchor for each class from the memory bank based on the finding tha t the cluster centers tend to be located in dense regions. In this technique an inter-contrast loss is derived from the perturbed unlabeled point cloud pairs wh ile an intra-contrast loss is derived from a single unlabeled point cloud. The d erived losses could enhance the discriminability of the features and implicitly constrain the semantic consistency between the perturbed unlabeled point cloud p airs. In addition we propose a dual-space hardness sampling strategy to pay more attention to the hard samples located in sparse regions of both the geometric s pace and feature space by reweighting the point-wise intra-contrast loss. Experi mental results on both indoor-scene and outdoor-scene datasets demonstrate that the proposed method outperforms the comparative state-of-the-art semi-supervised methods.
****
Adaptive Softassign via Hadamard-Equipped Sinkhorn
Binrui Shen, Qiang Niu, Shengxin Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17638-17647
Softassign is a pivotal method in graph matching and other learning tasks. Many

softassign-based algorithms exhibit performance sensitivity to a parameter in th e softassign. However tuning the parameter is challenging and almost done empiri cally. This paper proposes an adaptive softassign method for graph matching by a nalyzing the relationship between the objective score and the parameter. This me thod can automatically tune the parameter based on a given error bound to guaran tee accuracy. The Hadamard-Equipped Sinkhorn formulas introduced in this study s ignificantly enhance the efficiency and stability of the adaptive softassign. Mo reover these formulas can also be used in optimal transport problems. The result ing adaptive softassign graph matching algorithm enjoys significantly higher acc uracy than previous state-of-the-art large graph matching algorithms while maint aining comparable efficiency.
****
Re-thinking Data Availability Attacks Against Deep Neural Networks
Bin Fang, Bo Li, Shuang Wu, Shouhong Ding, Ran Yi, Lizhuang Ma; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12215-12224
The unauthorized use of personal data for commercial purposes and the covert acq uisition of private data for training machine learning models continue to raise concerns. To address these issues researchers have proposed availability attacks that aim to render data unexploitable. However many availability attack methods can be easily disrupted by adversarial training. Although some robust methods c an resist adversarial training their protective effects are limited. In this pap er we re-examine the existing availability attack methods and propose a novel tw o-stage min-max-min optimization paradigm to generate robust unlearnable noise. The inner min stage is utilized to generate unlearnable noise while the outer mi n-max stage simulates the training process of the poisoned model. Additionally w e formulate the attack effects and use it to constrain the optimization objectiv e. Comprehensive experiments have revealed that the noise generated by our metho d can lead to a decline in test accuracy for adversarially trained poisoned mode ls by up to approximately 30% in comparison to SOTA methods.
****
ElasticDiffusion: Training-free Arbitrary Size Image Generation through Global-L ocal Content Separation
Moayed Haji-Ali, Guha Balakrishnan, Vicente Ordonez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6603-66 12
Diffusion models have revolutionized image generation in recent years yet they a re still limited to a few sizes and aspect ratios. We propose ElasticDiffusion a novel training-free decoding method that enables pretrained text-to-image diffu sion models to generate images with various sizes. ElasticDiffusion attempts to decouple the generation trajectory of a pretrained model into local and global s ignals. The local signal controls low-level pixel information and can be estimat ed on local patches while the global signal is used to maintain overall structur al consistency and is estimated with a reference image. We test our method on Ce lebA-HQ (faces) and LAION-COCO (objects/indoor/outdoor scenes). Our experiments and qualitative results show superior image coherence quality across aspect rati os compared to MultiDiffusion and the standard decoding strategy of Stable Diffu sion. Project Webpage: https://elasticdiffusion.github.io
****
Locally Adaptive Neural 3D Morphable Models
Michail Tarasiou, Rolandos Alexandros Potamias, Eimear O'Sullivan, Stylianos Plo umpis, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2024, pp. 1867-1876
We present the Locally Adaptive Morphable Model (LAMM) a highly flexible Auto-En coder (AE) framework for learning to generate and manipulate 3D meshes. We train our architecture following a simple self-supervised training scheme in which in put displacements over a set of sparse control vertices are used to overwrite th e encoded geometry in order to transform one training sample into another. Durin g inference our model produces a dense output that adheres locally to the specif ied sparse geometry while maintaining the overall appearance of the encoded obje

ct. This approach results in state-of-the-art performance in both disentangling manipulated geometry and 3D mesh reconstruction. To the best of our knowledge LA MM is the first end-to-end framework that enables direct local control of 3D ver tex geometry in a single forward pass. A very efficient computational graph allo ws our network to train with only a fraction of the memory required by previous methods and run faster during inference generating 12k vertex meshes at >60fps o n a single CPU thread. We further leverage local geometry control as a primitive for higher level editing operations and present a set of derivative capabilitie s such as swapping and sampling object parts. Code and pretrained models can be found at https://github.com/michaeltrs/LAMM.
****
ICON: Incremental CONfidence for Joint Pose and Radiance Field Optimization
Weiyao Wang, Pierre Gleize, Hao Tang, Xingyu Chen, Kevin J Liang, Matt Feiszli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2024, pp. 5406-5417
Neural Radiance Fields (NeRF) exhibit remarkable performance for Novel View Synt hesis (NVS) given a set of 2D images. However NeRF training requires accurate ca mera pose for each input view typically obtained by Structure-from-Motion (SfM) pipelines. Recent works have attempted to relax this constraint but they still o ften rely on decent initial poses which they can refine. Here we aim at removing the requirement for pose initialization. We present Incremental CONfidence (ICO N) an optimization procedure for training NeRFs from 2D video frames. ICON only assumes smooth camera motion to estimate initial guess for poses. Further ICON i ntroduces "confidence": an adaptive measure of model quality used to dynamically reweight gradients. ICON relies on high-confidence poses to learn NeRF and high -confidence 3D structure (as encoded by NeRF) to learn poses. We show that ICON without prior pose initialization achieves superior performance in both CO3D and HO3D versus methods which use SfM pose.
****
Learned Scanpaths Aid Blind Panoramic Video Quality Assessment
Kanglong Fan, Wen Wen, Mu Li, Yifan Peng, Kede Ma; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2599-2608
Panoramic videos have the advantage of providing an immersive and interactive vi ewing experience. Nevertheless their spherical nature gives rise to various and uncertain user viewing behaviors which poses significant challenges for panorami c video quality assessment (PVQA). In this work we propose an end-to-end optimiz ed blind PVQA method with explicit modeling of user viewing patterns through vis ual scanpaths. Our method consists of two modules: a scanpath generator and a qu ality assessor. The scanpath generator is initially trained to predict future sc anpaths by minimizing their expected code length and then jointly optimized with the quality assessor for quality prediction. Our blind PVQA method enables dire ct quality assessment of panoramic images by treating them as videos composed of identical frames. Experiments on three public panoramic image and video quality datasets encompassing both synthetic and authentic distortions validate the sup eriority of our blind PVQA model over existing methods.
****
FineSports: A Multi-person Hierarchical Sports Video Dataset for Fine-grained Ac tion Understanding
Jinglin Xu, Guohao Zhao, Sibo Yin, Wenhao Zhou, Yuxin Peng; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21773-21782
Fine-grained action analysis in multi-person sports is complex due to athletes' quick movements and intense physical confrontations which result in severe visua l obstructions in most scenes. In addition accessible multi-person sports video datasets lack fine-grained action annotations in both space and time adding to t he difficulty in fine-grained action analysis. To this end we construct a new mu lti-person basketball sports video dataset named FineSports which contains fine-grained semantic and spatial-temporal annotations on 10000 NBA game videos cover ing 52 fine-grained action types 16000 action instances and 123000 spatial-tempo ral bounding boxes. We also propose a new prompt-driven spatial-temporal action

location approach called PoSTAL composed of a prompt-driven target action encode
r (PTA) and an action tube-specific detector (ATD) to directly generate target a
ction tubes with fine-grained action types without any off-line proposal generat
ion. Extensive experiments on the FineSports dataset demonstrate that PoSTAL out
performs state-of-the-art methods. Data and code are available at https://github
.com/PKU-ICST-MIPL/FineSports_CVPR2024.
****
SHiNe: Semantic Hierarchy Nexus for Open-vocabulary Object Detection
Mingxuan Liu, Tyler L. Hayes, Elisa Ricci, Gabriela Csurka, Riccardo Volpi; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 16634-16644
Open-vocabulary object detection (OvOD) has transformed detection into a languag
e-guided task empowering users to freely define their class vocabularies of inte
rest during inference. However our initial investigation indicates that existing
 OvOD detectors exhibit significant variability when dealing with vocabularies a
cross various semantic granularities posing a concern for real-world deployment.
 To this end we introduce Semantic Hierarchy Nexus (SHiNe) a novel classifier th
at uses semantic knowledge from class hierarchies. It runs offline in three step
s: i) it retrieves relevant super-/sub-categories from a hierarchy for each targ
et class; ii) it integrates these categories into hierarchy-aware sentences; iii
) it fuses these sentence embeddings to generate the nexus classifier vector. Ou
r evaluation on various detection benchmarks demonstrates that SHiNe enhances ro
bustness across diverse vocabulary granularities achieving up to +31.9% mAP50 wi
th ground truth hierarchies while retaining improvements using hierarchies gener
ated by large language models. Moreover when applied to open-vocabulary classifi
cation on ImageNet-1k SHiNe improves the CLIP zero-shot baseline by +2.8% accura
cy. SHiNe is training-free and can be seamlessly integrated with any off-the-she
lf OvOD detector without incurring additional computational overhead during infe
rence. The code is open source.
****
TI2V-Zero: Zero-Shot Image Conditioning for Text-to-Video Diffusion Models
Haomiao Ni, Bernhard Egger, Suhas Lohit, Anoop Cherian, Ye Wang, Toshiaki Koike-
Akino, Sharon X. Huang, Tim K. Marks; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9015-9025
Text-conditioned image-to-video generation (TI2V) aims to synthesize a realistic
 video starting from a given image (e.g. a woman's photo) and a text description
 (e.g. "a woman is drinking water."). Existing TI2V frameworks often require cos
tly training on video-text datasets and specific model designs for text and imag
e conditioning. In this paper we propose TI2V-Zero a zero-shot tuning-free metho
d that empowers a pretrained text-to-video (T2V) diffusion model to be condition
ed on a provided image enabling TI2V generation without any optimization fine-tu
ning or introducing external modules. Our approach leverages a pretrained T2V di
ffusion foundation model as the generative prior. To guide video generation with
 the additional image input we propose a "repeat-and-slide" strategy that modula
tes the reverse denoising process allowing the frozen diffusion model to synthes
ize a video frame-by-frame starting from the provided image. To ensure temporal
continuity we employ a DDPM inversion strategy to initialize Gaussian noise for
each newly synthesized frame and a resampling technique to help preserve visual
details. We conduct comprehensive experiments on both domain-specific and open-d
omain datasets where TI2V-Zero consistently outperforms a recent open-domain TI2
V model. Furthermore we show that TI2V-Zero can seamlessly extend to other tasks
 such as video infilling and prediction when provided with more images. Its auto
regressive design also supports long video generation.
****
Ranking Distillation for Open-Ended Video Question Answering with Insufficient L
abels
Tianming Liang, Chaolei Tan, Beihao Xia, Wei-Shi Zheng, Jian-Fang Hu; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 13161-13170
This paper focuses on open-ended video question answering which aims to find the

correct answers from a large answer set in response to a video-related question. This is essentially a multi-label classification task since a question may have multiple answers. However due to annotation costs the labels in existing benchmarks are always extremely insufficient typically one answer per question. As a result existing works tend to directly treat all the unlabeled answers as negative labels leading to limited ability for generalization. In this work we introduce a simple yet effective ranking distillation framework (RADI) to mitigate this problem without additional manual annotation. RADI employs a teacher model trained with incomplete labels to generate rankings for potential answers which contain rich knowledge about label priority as well as label-associated visual cues thereby enriching the insufficient labeling information. To avoid overconfidence in the imperfect teacher model we further present two robust and parameter-free ranking distillation approaches: a pairwise approach which introduces adaptive soft margins to dynamically refine the optimization constraints on various pairwise rankings and a listwise approach which adopts sampling-based partial listwise learning to resist the bias in teacher ranking. Extensive experiments on five popular benchmarks consistently show that both our pairwise and listwise RADIs outperform state-of-the-art methods. Further analysis demonstrates the effectiveness of our methods on the insufficient labeling problem.
****
GARField: Group Anything with Radiance Fields
Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, Angjoo Kanazawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21530-21539
Grouping is inherently ambiguous due to the multiple levels of granularity in which one can decompose a scene --- should the wheels of an excavator be considered separate or part of the whole? We propose Group Anything with Radiance Fields (GARField) an approach for decomposing 3D scenes into a hierarchy of semantically meaningful groups from posed image inputs. To do this we embrace group ambiguity through physical scale: by optimizing a scale-conditioned 3D affinity feature field a point in the world can belong to different groups of different sizes. We optimize this field from a set of 2D masks provided by Segment Anything (SAM) in a way that respects coarse-to-fine hierarchy using scale to consistently fuse conflicting masks from different viewpoints. From this field we can derive a hierarchy of possible groupings via automatic tree construction or user interaction. We evaluate GARField on a variety of in-the-wild scenes and find it effectively extracts groups at many levels: clusters of objects objects and various subparts. GARField inherently represents multi-view consistent groupings and produces higher fidelity groups than the input SAM masks. GARField's hierarchical grouping could have exciting downstream applications such as 3D asset extraction or dynamic scene understanding. Project site: https://www.garfield.studio/
****
Depth-Aware Concealed Crop Detection in Dense Agricultural Scenes
Liqiong Wang, Jinyu Yang, Yanfu Zhang, Fangyi Wang, Feng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17201-17211
Concealed Object Detection (COD) aims to identify objects visually embedded in their background. Existing COD datasets and methods predominantly focus on animals or humans ignoring the agricultural domain which often contains numerous small and concealed crops with severe occlusions. In this paper we introduce Concealed Crop Detection (CCD) which extends classic COD to agricultural domains. Experimental study shows that unimodal data provides insufficient information for CCD. To address this gap we first collect a large-scale RGB-D dataset ACOD-12K containing high-resolution crop images and depth maps. Then we propose a foundational framework named Recurrent Iterative Segmentation Network (RISNet). To tackle the challenge of dense objects we employ multi-scale receptive fields to capture objects of varying sizes thus enhancing the detection performance for dense objects. By fusing depth features our method can acquire spatial information about concealed objects to mitigate disturbances caused by intricate backgrounds and occlusions. Furthermore our model adopts a multi-stage iterative approach using pre

dictions from each stage as gate attention to reinforce position information the reby improving the detection accuracy for small objects. Extensive experimental results demonstrate that our RISNet achieves new state-of-the-art performance on both newly proposed CCD and classic COD tasks. All resources will be available at https://github.com/Kki2Eve/RISNet.
****
Learning Equi-angular Representations for Online Continual Learning
Minhyuk Seo, Hyunseo Koh, Wonje Jeung, Minjae Lee, San Kim, Hankook Lee, Sungjun Cho, Sungik Choi, Hyunwoo Kim, Jonghyun Choi; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23933-23942
Online continual learning suffers from an underfitted solution due to insufficie nt training for prompt model updates (e.g. single-epoch training). To address th e challenge we propose an efficient online continual learning method using the n eural collapse phenomenon. In particular we induce neural collapse to form a sim plex equiangular tight frame (ETF) structure in the representation space so that the continuously learned model with a single epoch can better fit to the stream ed data by proposing preparatory data training and residual correction in the re presentation space. With an extensive set of empirical validations using CIFAR-1 0/100 TinyImageNet ImageNet-200 and ImageNet-1K we show that our proposed method outperforms state-of-the-art methods by a noticeable margin in various online c ontinual learning scenarios such as disjoint and Gaussian scheduled continuous (i.e. boundary-free) data setups.
****
iToF-flow-based High Frame Rate Depth Imaging
Yu Meng, Zhou Xue, Xu Chang, Xuemei Hu, Tao Yue; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4929-4938
iToF is a prevalent cost-effective technology for 3D perception. While its relia nce on multi-measurement commonly leads to reduced performance in dynamic enviro nments. Based on the analysis of the physical iToF imaging process we propose th e iToF flow composed of crossmode transformation and uni-mode photometric correc tion to model the variation of measurements caused by different measurement mode s and 3D motion respectively. We propose a local linear transform (LLT) based cr oss-mode transfer module (LCTM) for mode-varying and pixel shift compensation of cross-mode flow and uni-mode photometric correct module (UPCM) for estimating t he depth-wise motion caused photometric residual of uni-mode flow. The iToF flow -based depth extraction network is proposed which could facilitate the estimatio n of the 4-phase measurements at each individual time for high framerate and acc urate depth estimation. Extensive experiments including both simulation and real -world experiments are conducted to demonstrate the effectiveness of the propose d methods. Compared with the SOTA method our approach reduces the computation ti me by 75% while improving the performance by 38%. The code and database are avai lable at https://github.com/ComputationalPerceptionLab/iToF_flow.
****
Solving the Catastrophic Forgetting Problem in Generalized Category Discovery
Xinzi Cao, Xiawu Zheng, Guanhong Wang, Weijiang Yu, Yunhang Shen, Ke Li, Yutong Lu, Yonghong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16880-16889
Generalized Category Discovery (GCD) aims to identify a mix of known and novel c ategories within unlabeled data sets providing a more realistic setting for imag e recognition. Essentially GCD needs to remember existing patterns thoroughly to recognize novel categories. Recent state-of-the-art method SimGCD transfers the knowledge from known-class data to the learning of novel classes through debias ed learning. However some patterns are catastrophically forgot during adaptation and thus lead to poor performance in novel categories classification. To addres s this issue we propose a novel learning approach LegoGCD which is seamlessly in tegrated into previous methods to enhance the discrimination of novel classes wh ile maintaining performance on previously encountered known classes. Specificall y we design two types of techniques termed as \underline L ocal \underline E ntr opy Re\underline g ularization (LER) and Dual-views Kullback-Leibler divergence c\underline o nstraint (DKL). The LER optimizes the distribution of potential kn

own class samples in unlabeled data thus ensuring the preservation of knowledge related to known categories while learning novel classes. Meanwhile DKL introduces Kullback-Leibler divergence to encourage the model to produce a similar prediction distribution of two view samples from the same image. In this way it successfully avoids mismatched prediction and generates more reliable potential known class samples simultaneously. Extensive experiments validate that the proposed LegoGCD effectively addresses the known category forgetting issue across all datasets e.g. delivering a 7.74% and 2.51% accuracy boost on known and novel classes in CUB respectively. Our code is available at: https://github.com/Cliffia123/LegoGCD.
****
Data-Efficient Unsupervised Interpolation Without Any Intermediate Frame for 4D Medical Images
JungEun Kim, Hangyul Yoon, Geondo Park, Kyungsu Kim, Eunho Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11353-11364
4D medical images which represent 3D images with temporal information are crucial in clinical practice for capturing dynamic changes and monitoring long-term disease progression. However acquiring 4D medical images poses challenges due to factors such as radiation exposure and imaging duration necessitating a balance between achieving high temporal resolution and minimizing adverse effects. Given these circumstances not only is data acquisition challenging but increasing the frame rate for each dataset also proves difficult. To address this challenge this paper proposes a simple yet effective Unsupervised Volumetric Interpolation framework UVI-Net. This framework facilitates temporal interpolation without the need for any intermediate frames distinguishing it from the majority of other existing unsupervised methods. Experiments on benchmark datasets demonstrate significant improvements across diverse evaluation metrics compared to unsupervised and supervised baselines. Remarkably our approach achieves this superior performance even when trained with a dataset as small as one highlighting its exceptional robustness and efficiency in scenarios with sparse supervision. This positions UVI-Net as a compelling alternative for 4D medical imaging particularly in settings where data availability is limited. The source code is available at https://github.com/jungeun122333/UVI-Net.
****
POCE: Primal Policy Optimization with Conservative Estimation for Multi-constraint Offline Reinforcement Learning
Jiayi Guan, Li Shen, Ao Zhou, Lusong Li, Han Hu, Xiaodong He, Guang Chen, Changjun Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26243-26253
Multi-constraint offline reinforcement learning (RL) promises to learn policies that satisfy both cumulative and state-wise costs from offline datasets. This arrangement provides an effective approach for the widespread application of RL in high-risk scenarios where both cumulative and state-wise costs need to be considered simultaneously. However previously constrained offline RL algorithms are primarily designed to handle single-constraint problems related to cumulative cost which faces challenges when addressing multi-constraint tasks that involve both cumulative and state-wise costs. In this work we propose a novel Primal policy Optimization with Conservative Estimation algorithm (POCE) to address the problem of multi-constraint offline RL. Concretely we reframe the objective of multi-constraint offline RL by introducing the concept of Maximum Markov Decision Processes (MMDP). Subsequently we present a primal policy optimization algorithm to confront the multi-constraint problems which improves the stability and convergence speed of model training. Furthermore we propose a conditional Bellman operator to estimate cumulative and state-wise Q-values reducing the extrapolation error caused by out-of-distribution (OOD) actions. Finally extensive experiments demonstrate that the POCE algorithm achieves competitive performance across multiple experimental tasks particularly outperforming baseline algorithms in terms of safety. Our code is available at \href https://github.com/guanjiayi/poce github.POCE .

****
Learning the 3D Fauna of the Web
Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9752-9762
Learning 3D models of all animals in nature requires massively scaling up existing solutions. With this ultimate goal in mind we develop 3D-Fauna an approach that learns a pan-category deformable 3D animal model for more than 100 animal species jointly. One crucial bottleneck of modeling animals is the limited availability of training data which we overcome by learning our model from 2D Internet images. We show that prior approaches which are category-specific fail to generalize to rare species with limited training images. We address this challenge by introducing the Semantic Bank of Skinned Models (SBSM) which automatically discovers a small set of base animal shapes by combining geometric inductive priors with semantic knowledge implicitly captured by an off-the-shelf self-supervised feature extractor. To train such a model we also contribute a new large-scale dataset of diverse animal species. At inference time given a single image of any quadruped animal our model reconstructs an articulated 3D mesh in a feed-forward manner in seconds.
****
Masked Spatial Propagation Network for Sparsity-Adaptive Depth Refinement
Jinyoung Jun, Jae-Han Lee, Chang-Su Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19768-19778
The main function of depth completion is to compensate for an insufficient and unpredictable number of sparse depth measurements of hardware sensors. However existing research on depth completion assumes that the sparsity --- the number of points or LiDAR lines --- is fixed for training and testing. Hence the completion performance drops severely when the number of sparse depths changes significantly. To address this issue we propose the sparsity-adaptive depth refinement (SDR) framework which refines monocular depth estimates using sparse depth points. For SDR we propose the masked spatial propagation network (MSPN) to perform SDR with a varying number of sparse depths effectively by gradually propagating sparse depth information throughout the entire depth map. Experimental results demonstrate that MPSN achieves state-of-the-art performance on both SDR and conventional depth completion scenarios.
****
LISA: Reasoning Segmentation via Large Language Model
Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9579-9589
Although perception systems have made remarkable advancements in recent years they still rely on explicit human instruction or pre-defined categories to identify the target objects before executing visual recognition tasks. Such systems cannot actively reason and comprehend implicit user intention. In this work we propose a new segmentation task --- reasoning segmentation. The task is designed to output a segmentation mask given a complex and implicit query text. Furthermore we establish a benchmark comprising over one thousand image-instruction-mask data samples incorporating intricate reasoning and world knowledge for evaluation purposes. Finally we present LISA: large Language Instructed Segmentation Assistant which inherits the language generation capabilities of multimodal Large Language Models (LLMs) while also possessing the ability to produce segmentation masks. We expand the original vocabulary with a <SEG> token and propose the embedding-as-mask paradigm to unlock the segmentation capability. Remarkably LISA can handle cases involving complex reasoning and world knowledge. Also it demonstrates robust zero-shot capability when trained exclusively on reasoning-free datasets. In addition fine-tuning the model with merely 239 reasoning segmentation data samples results in further performance enhancement. Both quantitative and qualitative experiments show our method effectively unlocks new reasoning segmentation capabilities for multimodal LLMs. Code models and data are available at github.com/dvlab-research/LISA.

****
Relightful Harmonization: Lighting-aware Portrait Background Replacement
Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, He Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6452-6462
Portrait harmonization aims to composite a subject into a new background adjusting its lighting and color to ensure harmony with the background scene. Existing harmonization techniques often only focus on adjusting the global color and brightness of the foreground and ignore crucial illumination cues from the background such as apparent lighting direction leading to unrealistic compositions. We introduce Relightful Harmonization a lighting-aware diffusion model designed to seamlessly harmonize sophisticated lighting effect for the foreground portrait using any background image. Our approach unfolds in three stages. First we introduce a lighting representation module that allows our diffusion model to encode lighting information from target image background. Second we introduce an alignment network that aligns lighting features learned from image background with lighting features learned from panorama environment maps which is a complete representation for scene illumination. Last to further boost the photorealism of the proposed method we introduce a novel data simulation pipeline that generates synthetic training pairs from a diverse range of natural images which are used to refine the model. Our method outperforms existing benchmarks in visual fidelity and lighting coherence showing superior generalization in real-world testing scenarios highlighting its versatility and practicality.
****
Bridging the Gap: A Unified Video Comprehension Framework for Moment Retrieval and Highlight Detection
Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, Xiu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18709-18719
Video Moment Retrieval (MR) and Highlight Detection (HD) have attracted significant attention due to the growing demand for video analysis. Recent approaches treat MR and HD as similar video grounding problems and address them together with transformer-based architecture. However we observe that the emphasis of MR and HD differs with one necessitating the perception of local relationships and the other prioritizing the understanding of global contexts. Consequently the lack of task-specific design will inevitably lead to limitations in associating the intrinsic specialty of two tasks. To tackle the issue we propose a Unified Video COMprehension framework (UVCOM) to bridge the gap and jointly solve MR and HD effectively. By performing progressive integration on intra and inter-modality across multi-granularity UVCOM achieves the comprehensive understanding in processing a video. Moreover we present multi-aspect contrastive learning to consolidate the local relation modeling and global knowledge accumulation via well aligned multi-modal space. Extensive experiments on QVHighlights Charades-STA TACoS YouTube Highlights and TVSum datasets demonstrate the effectiveness and rationality of UVCOM which outperforms the state-of-the-art methods by a remarkable margin.
****
MuseChat: A Conversational Music Recommendation System for Videos
Zhikang Dong, Xiulong Liu, Bin Chen, Pawel Polak, Peng Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12775-12785
Music recommendation for videos attracts growing interest in multi-modal research. However existing systems focus primarily on content compatibility often ignoring the users' preferences. Their inability to interact with users for further refinements or to provide explanations leads to a less satisfying experience. We address these issues with MuseChat a first-of-its-kind dialogue-based recommendation system that personalizes music suggestions for videos. Our system consists of two key functionalities with associated modules: recommendation and reasoning. The recommendation module takes a video along with optional information including previous suggested music and user's preference as inputs and retrieves an appropriate music matching the context. The reasoning module equipped with the pow

er of Large Language Model (Vicuna-7B) and extended to multi-modal inputs is able to provide reasonable explanation for the recommended music. To evaluate the effectiveness of MuseChat we build a large-scale dataset conversational music recommendation for videos that simulates a two-turn interaction between a user and a recommender based on accurate music track information. Experiment results show that MuseChat achieves significant improvements over existing video-based music retrieval methods as well as offers strong interpretability and interactability. The dataset of this work is available at https://dongzhikang.github.io/musechat.
****
Mitigating Motion Blur in Neural Radiance Fields with Events and Frames
Marco Cannici, Davide Scaramuzza; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9286-9296
Neural Radiance Fields (NeRFs) have shown great potential in novel view synthesis. However they struggle to render sharp images when the data used for training is affected by motion blur. On the other hand event cameras excel in dynamic scenes as they measure brightness changes with microsecond resolution and are thus only marginally affected by blur. Recent methods attempt to enhance NeRF reconstructions under camera motion by fusing frames and events. However they face challenges in recovering accurate color content or constrain the NeRF to a set of predefined camera poses harming reconstruction quality in challenging conditions. This paper proposes a novel formulation addressing these issues by leveraging both model- and learning-based modules. We explicitly model the blur formation process exploiting the event double integral as an additional model-based prior. Additionally we model the event-pixel response using an end-to-end learnable response function allowing our method to adapt to non-idealities in the real event-camera sensor. We show on synthetic and real data that the proposed approach outperforms existing deblur NeRFs that use only frames as well as those that combine frames and events by +6.13dB and +2.48dB respectively.
****
C3Net: Compound Conditioned ControlNet for Multimodal Content Generation
Juntao Zhang, Yuehuai Liu, Yu-Wing Tai, Chi-Keung Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26886-26895
We present Compound Conditioned ControlNet C3Net a novel generative neural architecture taking conditions from multiple modalities and synthesizing multimodal contents simultaneously (e.g. image text audio). C3Net adapts the ControlNet architecture to jointly train and make inferences on a production-ready diffusion model and its trainable copies. Specifically C3Net first aligns the conditions from multi-modalities to the same semantic latent space using modality-specific encoders based on contrastive training. Then it generates multimodal outputs based on the aligned latent space whose semantic information is combined using a ControlNet-like architecture called Control C3-UNet. Correspondingly with this system design our model offers an improved solution for joint-modality generation through learning and explaining multimodal conditions involving more than just linear interpolation within the latent space. Meanwhile as we align conditions to a unified latent space C3Net only requires one trainable Control C3-UNet to work on multimodal semantic information. Furthermore our model employs unimodal pretraining on the condition alignment stage outperforming the non-pretrained alignment even on relatively scarce training data and thus demonstrating high-quality compound condition generation. We contribute the first high-quality tri-modal validation set to validate quantitatively that C3Net outperforms or is on par with the first and contemporary state-of-the-art multimodal generation. Our codes and tri-modal dataset will be released.
****
Device-Wise Federated Network Pruning
Shangqian Gao, Junyi Li, Zeyu Zhang, Yanfu Zhang, Weidong Cai, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12342-12352
Neural network pruning particularly channel pruning is a widely used technique f

or compressing deep learning models to enable their deployment on edge devices with limited resources. Typically redundant weights or structures are removed to achieve the target resource budget. Although data-driven pruning approaches have proven to be more effective they cannot be directly applied to federated learning (FL) which has emerged as a popular technique in edge computing applications because of distributed and confidential datasets. In response to this challenge we design a new network pruning method for FL. We propose device-wise sub-networks for each device assuming that the data distribution is similar within each device. These sub-networks are generated through sub-network embeddings and a hypernetwork. To further minimize memory usage and communication costs we permanently prune the full model to remove weights that are not useful for all devices. During the FL process we simultaneously train the device-wise sub-networks and the base sub-network to facilitate the pruning process. We then finetune the pruned model with device-wise sub-networks to regain performance. Moreover we provided the theoretical guarantee of convergence for our method. Our method achieves better performance and resource trade-off than other well-established network pruning baselines as demonstrated through extensive experiments on CIFAR-10 CIFAR-100 and TinyImageNet.
****
Adapt Before Comparison: A New Perspective on Cross-Domain Few-Shot Segmentation
Jonas Herzog; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23605-23615
Few-shot segmentation performance declines substantially when facing images from a domain different than the training domain effectively limiting real-world use cases. To alleviate this recently cross-domain few-shot segmentation (CD-FSS) has emerged. Works that address this task mainly attempted to learn segmentation on a source domain in a manner that generalizes across domains. Surprisingly we can outperform these approaches while eliminating the training stage and removing their main segmentation network. We show test-time task-adaption is the key for successful CD-FSS instead. Task-adaption is achieved by appending small networks to the feature pyramid of a conventionally classification-pretrained backbone. To avoid overfitting to the few labeled samples in supervised fine-tuning consistency across augmented views of input images serves as guidance while learning the parameters of the attached layers. Despite our self-restriction not to use any images other than the few labeled samples at test time we achieve new state-of-the-art performance in CD-FSS evidencing the need to rethink approaches for the task. Code is available at https://github.com/Vision-Kek/ABCDFSS.
****
TokenHMR: Advancing Human Mesh Recovery with a Tokenized Pose Representation
Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1323-1333
We address the problem of regressing 3D human pose and shape from a single image with a focus on 3D accuracy. The current best methods leverage large datasets of 3D pseudo-ground-truth (p-GT) and 2D keypoints leading to robust performance. With such methods however we observe a paradoxical decline in 3D pose accuracy with increasing 2D accuracy. This is caused by biases in the p-GT and the use of an approximate camera projection model. We quantify the error induced by current camera models and show that fitting 2D keypoints and p-GT accurately causes incorrect 3D poses. Our analysis defines the invalid distances within which minimizing 2D and p-GT losses is detrimental. We use this to formulate a new loss "Threshold-Adaptive Loss Scaling" (TALS) that penalizes gross 2D and p-GT errors but not smaller ones. With such a loss there are many 3D poses that could equally explain the 2D evidence. To reduce this ambiguity we need a prior over valid human poses but such priors can introduce unwanted bias. To address this we exploit a tokenized representation of human pose and reformulate the problem as token prediction. This restricts the estimated poses to the space of valid poses effectively improving robustness to occlusion. Extensive experiments on the EMDB and 3DPW datasets show that our reformulated loss and tokenization allows us to train on in-the-wild data while improving 3D accuracy over the state-of-the-art. Our mo

dels and code are available for research at https://tokenhmr.is.tue.mpg.de.
****
MoReVQA: Exploring Modular Reasoning Models for Video Question Answering
Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13235-13245
This paper addresses the task of video question answering (videoQA) via a decomposed multi-stage modular reasoning framework. Previous modular methods have shown promise with a single planning stage ungrounded in visual content. However through a simple and effective baseline we find that such systems can lead to brittle behavior in practice for challenging videoQA settings. Thus unlike traditional single-stage planning methods we propose a multi-stage system consisting of an event parser a grounding stage and a final reasoning stage in conjunction with an external memory. All stages are training-free and performed using few-shot prompting of large models creating interpretable intermediate outputs at each stage. By decomposing the underlying planning and task complexity our method MoReVQA improves over prior work on standard videoQA benchmarks (NExT-QA iVQA EgoSchema and ActivityNet-QA) with state-of-the-art results and extensions to related tasks (grounded videoQA paragraph captioning).
****
Low-Rank Rescaled Vision Transformer Fine-Tuning: A Residual Design Approach
Wei Dong, Xing Zhang, Bihui Chen, Dawei Yan, Zhijun Lin, Qingsen Yan, Peng Wang, Yang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16101-16110
Parameter-efficient fine-tuning for pre-trained Vision Transformers aims to adeptly tailor a model to downstream tasks by learning a minimal set of new adaptation parameters while preserving the frozen majority of pre-trained parameters. Striking a balance between retaining the generalizable representation capacity of the pre-trained model and acquiring task-specific features poses a key challenge. Currently there is a lack of focus on guiding this delicate trade-off. In this study we approach the problem from the perspective of Singular Value Decomposition (SVD) of pre-trained parameter matrices providing insights into the tuning dynamics of existing methods. Building upon this understanding we propose a Residual-based Low-Rank Rescaling (RLRR) fine-tuning strategy. This strategy not only enhances flexibility in parameter tuning but also ensures that new parameters do not deviate excessively from the pre-trained model through a residual design. Extensive experiments demonstrate that our method achieves competitive performance across various downstream image classification tasks all while maintaining comparable new parameters. We believe this work takes a step forward in offering a unified perspective for interpreting existing methods and serves as motivation for the development of new approaches that move closer to effectively considering the crucial trade-off mentioned above. Our code is available at https://github.com/zstarN70/RLRR.git.
****
FaceCom: Towards High-fidelity 3D Facial Shape Completion via Optimization and Inpainting Guidance
Yinglong Li, Hongyu Wu, Xiaogang Wang, Qingzhao Qin, Yijiao Zhao, Yong Wang, Aimin Hao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2177-2186
We propose FaceCom a method for 3D facial shape completion which delivers high-fidelity results for incomplete facial inputs of arbitrary forms. Unlike end-to-end shape completion methods based on point clouds or voxels our approach relies on a mesh-based generative network that is easy to optimize enabling it to handle shape completion for irregular facial scans. We first train a shape generator on a mixed 3D facial dataset containing 2405 identities. Based on the incomplete facial input we fit complete faces using an optimization approach under image inpainting guidance. The completion results are refined through a post-processing step. FaceCom demonstrates the ability to effectively and naturally complete facial scan data with varying missing regions and degrees of missing areas. Our method can be used in medical prosthetic fabrication and the registration of defic

ient scanning data. Our experimental results demonstrate that FaceCom achieves exceptional performance in fitting and shape completion tasks.
****
Distribution-aware Knowledge Prototyping for Non-exemplar Lifelong Person Re-identification
Kunlun Xu, Xu Zou, Yuxin Peng, Jiahuan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16604-16613
Lifelong person re-identification (LReID) suffers from the catastrophic forgetting problem when learning from non-stationary data. Existing exemplar-based and knowledge distillation-based LReID methods encounter data privacy and limited acquisition capacity respectively. In this paper we instead introduce the prototype which is under-investigated in LReID to better balance knowledge forgetting and acquisition. Existing prototype-based works primarily focus on the classification task where the prototypes are set as discrete points or statistical distributions. However they either discard the distribution information or omit instance-level diversity which are crucial fine-grained clues for LReID. To address the above problems we propose Distribution-aware Knowledge Prototyping (DKP) where the instance-level diversity of each sample is modeled to transfer comprehensive fine-grained knowledge for prototyping and facilitating LReID learning. Specifically an Instance-level Distribution Modeling network is proposed to capture the local diversity of each instance. Then the Distribution-oriented Prototype Generation algorithm transforms the instance-level diversity into identity-level distributions as prototypes which is further explored by the designed Prototype-based Knowledge Transfer module to enhance the knowledge anti-forgetting and acquisition capacity of the LReID model. Extensive experiments verify that our method achieves superior plasticity and stability balancing and outperforms existing LReID methods by 8.1%/9.1% average mAP/R@1 improvement. The code is available at https://github.com/zhoujiahuan1991/CVPR2024-DKP
****
LightOctree: Lightweight 3D Spatially-Coherent Indoor Lighting Estimation
Xuecan Wang, Shibang Xiao, Xiaohui Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4536-4545
We present a lightweight solution for estimating spatially-coherent indoor lighting from a single RGB image. Previous methods for estimating illumination using volumetric representations have overlooked the sparse distribution of light sources in space necessitating substantial memory and computational resources for achieving high-quality results. We introduce a unified voxel octree-based illumination estimation framework to produce 3D spatially-coherent lighting. Additionally a differentiable voxel octree cone tracing rendering layer is proposed to eliminate regular volumetric representation throughout the entire process and ensure the retention of features across different frequency domains. This reduction significantly decreases spatial usage and required floating-point operations without substantially compromising precision. Experimental results demonstrate that our approach achieves high-quality coherent estimation with minimal cost compared to previous methods.
****
Generating Enhanced Negatives for Training Language-Based Object Detectors
Shiyu Zhao, Long Zhao, Vijay Kumar B G, Yumin Suh, Dimitris N. Metaxas, Manmohan Chandraker, Samuel Schulter; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13592-13602
The recent progress in language-based open-vocabulary object detection can be largely attributed to finding better ways of leveraging large-scale data with free-form text annotations. Training such models with a discriminative objective function has proven successful but requires good positive and negative samples. However the free-form nature and the open vocabulary of object descriptions make the space of negatives extremely large. Prior works randomly sample negatives or use rule-based techniques to build them. In contrast we propose to leverage the vast knowledge built into modern generative models to automatically build negatives that are more relevant to the original data. Specifically we use large-language-models to generate negative text descriptions and text-to-image diffusion mod

els to also generate corresponding negative images. Our experimental analysis confirms the relevance of the generated negative data and its use in language-based detectors improves performance on two complex benchmarks. Code is available at https://github.com/xiaofeng94/Gen-Enhanced-Negs.
****
Insect-Foundation: A Foundation Model and Large-scale 1M Dataset for Visual Insect Understanding
Hoang-Quan Nguyen, Thanh-Dat Truong, Xuan Bac Nguyen, Ashley Dowling, Xin Li, Khoa Luu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21945-21955
In precision agriculture the detection and recognition of insects play an essential role in the ability of crops to grow healthy and produce a high-quality yield. The current machine vision model requires a large volume of data to achieve high performance. However there are approximately 5.5 million different insect species in the world. None of the existing insect datasets can cover even a fraction of them due to varying geographic locations and acquisition costs. In this paper we introduce a novel "Insect-1M" dataset a game-changing resource poised to revolutionize insect-related foundation model training. Covering a vast spectrum of insect species our dataset including 1 million images with dense identification labels of taxonomy hierarchy and insect descriptions offers a panoramic view of entomology enabling foundation models to comprehend visual and semantic information about insects like never before. Then to efficiently establish an Insect Foundation Model we develop a micro-feature self-supervised learning method with a Patch-wise Relevant Attention mechanism capable of discerning the subtle differences among insect images. In addition we introduce Description Consistency loss to improve micro-feature modeling via insect descriptions. Through our experiments we illustrate the effectiveness of our proposed approach in insect modeling and achieve State-of-the-Art performance on standard benchmarks of insect-related tasks. Our Insect Foundation Model and Dataset promise to empower the next generation of insect-related vision models bringing them closer to the ultimate goal of precision agriculture.
****
Data-Efficient Multimodal Fusion on a Single GPU
Noël Vouitsis, Zhaoyan Liu, Satya Krishna Gorti, Valentin Villecroze, Jesse C. Cresswell, Guangwei Yu, Gabriel Loaiza-Ganem, Maksims Volkovs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27239-27251
The goal of multimodal alignment is to learn a single latent space that is shared between multimodal inputs. The most powerful models in this space have been trained using massive datasets of paired inputs and large-scale computational resources making them prohibitively expensive to train in many practical scenarios. We surmise that existing unimodal encoders pre-trained on large amounts of unimodal data should provide an effective bootstrap to create multimodal models from unimodal ones at much lower costs. We therefore propose FuseMix a multimodal augmentation scheme that operates on the latent spaces of arbitrary pre-trained unimodal encoders. Using FuseMix for multimodal alignment we achieve competitive performance - and in certain cases outperform state-of-the art methods - in both image-text and audio-text retrieval with orders of magnitude less compute and data: for example we outperform CLIP on the Flickr30K text-to-image retrieval task with ?600x fewer GPU days and ?80x fewer image-text pairs. Additionally we show how our method can be applied to convert pre-trained text-to-image generative models into audio-to-image ones. Code is available at: https://github.com/layer6ai-labs/fusemix.
****
FedSelect: Personalized Federated Learning with Customized Selection of Parameters for Fine-Tuning
Rishub Tamirisa, Chulin Xie, Wenxuan Bao, Andy Zhou, Ron Arel, Aviv Shamsian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23985-23994
Standard federated learning approaches suffer when client data distributions hav

e sufficient heterogeneity. Recent methods addressed the client data heterogeneity issue via personalized federated learning (PFL) - a class of FL algorithms aiming to personalize learned global knowledge to better suit the clients' local data distributions. Existing PFL methods usually decouple global updates in deep neural networks by performing personalization on particular layers (i.e. classifier heads) and global aggregation for the rest of the network. However preselecting network layers for personalization may result in suboptimal storage of global knowledge. In this work we propose FedSelect a novel PFL algorithm inspired by the iterative subnetwork discovery procedure used for the Lottery Ticket Hypothesis. FedSelect incrementally expands subnetworks to personalize client parameters concurrently conducting global aggregations on the remaining parameters. This approach enables the personalization of both client parameters and subnetwork structure during the training process. Finally we show that FedSelect outperforms recent state-of-the-art PFL algorithms under challenging client data heterogeneity settings and demonstrates robustness to various real-world distributional shifts.
****
FaceLift: Semi-supervised 3D Facial Landmark Localization
David Ferman, Pablo Garrido, Gaurav Bharaj; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1781-1791
3D facial landmark localization has proven to be of particular use for applications such as face tracking 3D face modeling and image-based 3D face reconstruction. In the supervised learning case such methods usually rely on 3D landmark datasets derived from 3DMM-based registration that often lack spatial definition alignment as compared with that chosen by hand-labeled human consensus e.g. how are eyebrow landmarks defined? This creates a gap between landmark datasets generated via high-quality 2D human labels and 3DMMs and it ultimately limits their effectiveness. To address this issue we introduce a novel semi-supervised learning approach that learns 3D landmarks by directly lifting (visible) hand-labeled 2D landmarks and ensures better definition alignment without the need for 3D landmark datasets. To lift 2D landmarks to 3D we leverage 3D-aware GANs for better multi-view consistency learning and in-the-wild multi-frame videos for robust cross-generalization. Empirical experiments demonstrate that our method not only achieves better definition alignment between 2D-3D landmarks but also outperforms other supervised learning 3D landmark localization methods on both 3DMM labeled and photogrammetric ground truth evaluation datasets. Project Page: https://davidcferman.github.io/FaceLift
****
PSDPM: Prototype-based Secondary Discriminative Pixels Mining for Weakly Supervised Semantic Segmentation
Xinqiao Zhao, Ziqian Yang, Tianhong Dai, Bingfeng Zhang, Jimin Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3437-3446
Image-level Weakly Supervised Semantic Segmentation (WSSS) has received increasing attention due to its low annotation cost. Class Activation Mapping (CAM) generated through classifier weights in WSSS inevitably ignores certain useful cues while the CAM generated through class prototypes can alleviate that. However because of the different goals of image classification and semantic segmentation the class prototypes still focus on activating primary discriminative pixels learned from classification loss leading to incomplete CAM. In this paper we propose a plugand-play Prototype-based Secondary Discriminative Pixels Mining (PSDPM) framework for enabling class prototypes to activate more secondary discriminative pixels thus generating a more complete CAM. Specifically we introduce a Foreground Pixel Estimation Module (FPEM) for estimating potential foreground pixels based on the correlations between primary and secondary discriminative pixels and the semantic segmentation results of baseline methods. Then we enable WSSS model to learn discriminative features from secondary discriminative pixels through a consistency loss calculated between FPEM result and class-prototype CAM. Experimental results show that our PSDPM improves various baseline methods significantly and achieves new state-of-the-art performances on WSSS benchmarks. Codes are a

vailable at https://github.com/xinqiaozhao/PSDPM.
****
Bidirectional Multi-Scale Implicit Neural Representations for Image Deraining
Xiang Chen, Jinshan Pan, Jiangxin Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25627-25636
How to effectively explore multi-scale representations of rain streaks is important for image deraining. In contrast to existing Transformer-based methods that depend mostly on single-scale rain appearance we develop an end-to-end multi-scale Transformer that leverages the potentially useful features in various scales to facilitate high-quality image reconstruction. To better explore the common degradation representations from spatially-varying rain streaks we incorporate intra-scale implicit neural representations based on pixel coordinates with the degraded inputs in a closed-loop design enabling the learned features to facilitate rain removal and improve the robustness of the model in complex scenarios. To ensure richer collaborative representation from different scales we embed a simple yet effective inter-scale bidirectional feedback operation into our multi-scale Transformer by performing coarse-to-fine and fine-to-coarse information communication. Extensive experiments demonstrate that our approach named as NeRD-Rain performs favorably against the state-of-the-art ones on both synthetic and real-world benchmark datasets. The source code and trained models are available at https://github.com/cschenxiang/NeRD-Rain.
****
Frozen CLIP: A Strong Backbone for Weakly Supervised Semantic Segmentation
Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, Jimin Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3796-3806
Weakly supervised semantic segmentation has witnessed great achievements with image-level labels. Several recent approaches use the CLIP model to generate pseudo labels for training an individual segmentation model while there is no attempt to apply the CLIP model as the backbone to directly segment objects with image-level labels. In this paper we propose WeCLIP a CLIP-based single-stage pipeline for weakly supervised semantic segmentation. Specifically the frozen CLIP model is applied as the backbone for semantic feature extraction and a new decoder is designed to interpret extracted semantic features for final prediction. Meanwhile we utilize the above frozen backbone to generate pseudo labels for training the decoder. Such labels cannot be optimized during training. We then propose a refinement module (RFM) to rectify them dynamically. Our architecture enforces the proposed decoder and RFM to benefit from each other to boost the final performance. Extensive experiments show that our approach significantly outperforms other approaches with less training cost. Additionally our WeCLIP also obtains promising results for fully supervised settings. The code is available at https://github.com/zbf1991/WeCLIP.
****
FedAS: Bridging Inconsistency in Personalized Federated Learning
Xiyuan Yang, Wenke Huang, Mang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11986-11995
Personalized Federated Learning (PFL) is primarily designed to provide customized models for each client to better fit the non-iid distributed client data which is a inherent challenge in Federated Learning. However current PFL methods suffer from inconsistencies in both intra-client and inter-client levels: 1) The intra-client inconsistency stems from the asynchronous update strategy for personalized and shared parameters. In PFL clients update their shared parameters to communicate and learn from others while keeping personalized parts unchanged leading to poor coordination between these two components. 2) The Inter-client inconsistency arises from "stragglers" - inactive clients that communicate and train with the server less frequently. This results in their under-trained personalized models and impedes the collaborative training stage for other clients. In this paper we present a novel PFL framework named FedAS which uses Federated Parameter-Alignment and Client-Synchronization to overcome above challenges. Initially we enhance the localization of global parameters by infusing them with local insig

hts. We make the shared parts learn from previous model thereby increasing their
 local relevance and reducing the impact of parameter inconsistency. Furthermore
 we design a robust aggregation method to mitigate the impact of stragglers by p
reventing the incorporation of their under-trained knowledge into aggregated mod
el. Experimental results on Cifar10 and Cifar100 validate the effectiveness of o
ur FedAS in achieving better performance and robustness against data heterogenei
ty.
****
LAFS: Landmark-based Facial Self-supervised Learning for Face Recognition
Zhonglin Sun, Chen Feng, Ioannis Patras, Georgios Tzimiropoulos; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 1639-1649
In this work we focus on learning facial representations that can be adapted to
train effective face recognition models particularly in the absence of labels. F
irstly compared with existing labelled face datasets a vastly larger magnitude o
f unlabeled faces exists in the real world. We explore the learning strategy of
these unlabeled facial images through self-supervised pretraining to transfer ge
neralized face recognition performance. Moreover motivated by one recent finding
 that is the face saliency area is critical for face recognition in contrast to
utilizing random cropped blocks of images for constructing augmentations in pret
raining we utilize patches localized by extracted facial landmarks. This enables
 our method - namely Landmark-based Facial Self-supervised learning (LAFS) to le
arn key representation that is more critical for face recognition. We also incor
porate two landmark-specific augmentations which introduce more diversity of lan
dmark information to further regularize the learning. With learned landmark-base
d facial representations we further adapt the representation for face recognitio
n with regularization mitigating variations in landmark positions. Our method ac
hieves significant improvement over the state-of-the-art on multiple face recogn
ition benchmarks especially on more challenging few-shot scenarios. The code is
available at https://github.com/szlbiubiubiu/LAFS_CVPR2024
****
SED: A Simple Encoder-Decoder for Open-Vocabulary Semantic Segmentation
Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, Yanwei Pang; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 3426-3436
Open-vocabulary semantic segmentation strives to distinguish pixels into differe
nt semantic groups from an open set of categories. Most existing methods explore
 utilizing pre-trained vision-language models in which the key is to adopt the i
mage-level model for pixel-level segmentation task. In this paper we propose a s
imple encoder-decoder named SED for open-vocabulary semantic segmentation which
comprises a hierarchical encoder-based cost map generation and a gradual fusion
decoder with category early rejection. The hierarchical encoder-based cost map g
eneration employs hierarchical backbone instead of plain transformer to predict
pixel-level image-text cost map. Compared to plain transformer hierarchical back
bone better captures local spatial information and has linear computational comp
lexity with respect to input size. Our gradual fusion decoder employs a top-down
 structure to combine cost map and the feature maps of different backbone levels
 for segmentation. To accelerate inference speed we introduce a category early r
ejection scheme in the decoder that rejects many no-existing categories at the e
arly layer of decoder resulting in at most 4.7 times acceleration without accura
cy degradation. Experiments are performed on multiple open-vocabulary semantic s
egmentation datasets which demonstrates the efficacy of our SED method. When usi
ng ConvNeXt-B our SED method achieves mIoU score of 31.6% on ADE20K with 150 cat
egories at 82 millisecond (ms) per image on a single A6000. Our source code is a
vailable at https://github.com/xb534/SED.
****
GPLD3D: Latent Diffusion of 3D Shape Generative Models by Enforcing Geometric an
d Physical Priors
Yuan Dong, Qi Zuo, Xiaodong Gu, Weihao Yuan, Zhengyi Zhao, Zilong Dong, Liefeng
Bo, Qixing Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR), 2024, pp. 56-66
State-of-the-art man-made shape generative models usually adopt established generative models under a suitable implicit shape representation. A common theme is to perform distribution alignment which does not explicitly model important shape priors. As a result many synthetic shapes are not connected. Other synthetic shapes present problems of physical stability and geometric feasibility. This paper introduces a novel latent diffusion shape-generative model regularized by a quality checker that outputs a score of a latent code. The scoring function employs a learned function that provides a geometric feasibility score and a deterministic procedure to quantify a physical stability score. The key to our approach is a new diffusion procedure that combines the discrete empirical data distribution and a continuous distribution induced by the quality checker. We introduce a principled approach to determine the tradeoff parameters for learning the denoising network at different noise levels. Experimental results show that our approach outperforms state-of-the-art shape generations quantitatively and qualitatively on ShapeNet-v2.
****
Enhancing Quality of Compressed Images by Mitigating Enhancement Bias Towards Compression Domain
Qunliang Xing, Mai Xu, Shengxi Li, Xin Deng, Meisong Zheng, Huaida Liu, Ying Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25501-25511
Existing quality enhancement methods for compressed images focus on aligning the enhancement domain with the raw domain to yield realistic images. However these methods exhibit a pervasive enhancement bias towards the compression domain inadvertently regarding it as more realistic than the raw domain. This bias makes enhanced images closely resemble their compressed counterparts thus degrading their perceptual quality. In this paper we propose a simple yet effective method to mitigate this bias and enhance the quality of compressed images. Our method employs a conditional discriminator with the compressed image as a key condition and then incorporates a domain-divergence regularization to actively distance the enhancement domain from the compression domain. Through this dual strategy our method enables the discrimination against the compression domain and brings the enhancement domain closer to the raw domain. Comprehensive quality evaluations confirm the superiority of our method over other state-of-the-art methods without incurring inference overheads.
****
LangSplat: 3D Language Gaussian Splatting
Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, Hanspeter Pfister; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20051-20060
Humans live in a 3D world and commonly use natural language to interact with a 3D scene. Modeling a 3D language field to support open-ended language queries in 3D has gained increasing attention recently. This paper introduces LangSplat which constructs a 3D language field that enables precise and efficient open-vocabulary querying within 3D spaces. Unlike existing methods that ground CLIP language embeddings in a NeRF model LangSplat advances the field by utilizing a collection of 3D Gaussians each encoding language features distilled from CLIP to represent the language field. By employing a tile-based splatting technique for rendering language features we circumvent the costly rendering process inherent in NeRF. Instead of directly learning CLIP embeddings LangSplat first trains a scene-wise language autoencoder and then learns language features on the scene-specific latent space thereby alleviating substantial memory demands imposed by explicit modeling. Existing methods struggle with imprecise and vague 3D language fields which fail to discern clear boundaries between objects. We delve into this issue and propose to learn hierarchical semantics using SAM thereby eliminating the need for extensively querying the language field across various scales and the regularization of DINO features. Extensive experimental results show that LangSplat significantly outperforms the previous state-of-the-art method LERF by a large margin. Notably LangSplat is extremely efficient achieving a 199 x speedup co

mpared to LERF at the resolution of 1440 x 1080. We strongly recommend readers t
o check out our video results at https://langsplat.github.io/.
****
MoST: Multi-Modality Scene Tokenization for Motion Prediction
Norman Mu, Jingwei Ji, Zhenpei Yang, Nate Harada, Haotian Tang, Kan Chen, Charle
s R. Qi, Runzhou Ge, Kratarth Goel, Zoey Yang, Scott Ettinger, Rami Al-Rfou, Dra
gomir Anguelov, Yin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2024, pp. 14988-14999
Many existing motion prediction approaches rely on symbolic perception outputs t
o generate agent trajectories such as bounding boxes road graph information and
traffic lights. This symbolic representation is a high-level abstraction of the
real world which may render the motion prediction model vulnerable to perception
 errors (e.g. failures in detecting open-vocabulary obstacles) while missing sal
ient information from the scene context (e.g. poor road conditions). An alternat
ive paradigm is end-to-end learning from raw sensors. However this approach suff
ers from the lack of interpretability and requires significantly more training r
esources. In this work we propose tokenizing the visual world into a compact set
 of scene elements and then leveraging pre-trained image foundation models and L
iDAR neural networks to encode all the scene elements in an open-vocabulary mann
er. The image foundation model enables our scene tokens to encode the general kn
owledge of the open world while the LiDAR neural network encodes geometry inform
ation. Our proposed representation can efficiently encode the multi-frame multi-
modality observations with a few hundred tokens and is compatible with most tran
sformer-based architectures. To evaluate our method we have augmented Waymo Open
 Motion Dataset with camera embeddings. Experiments over Waymo Open Motion Datas
et show that our approach leads to significant performance improvements over the
 state-of-the-art.
****
PIGEON: Predicting Image Geolocations
Lukas Haas, Michal Skreta, Silas Alberti, Chelsea Finn; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1289
3-12902
Planet-scale image geolocalization remains a challenging problem due to the dive
rsity of images originating from anywhere in the world. Although approaches base
d on vision transformers have made significant progress in geolocalization accur
acy success in prior literature is constrained to narrow distributions of images
 of landmarks and performance has not generalized to unseen places. We present a
 new geolocalization system that combines semantic geocell creation multi-task c
ontrastive pretraining and a novel loss function. Additionally our work is the f
irst to perform retrieval over location clusters for guess refinements. We train
 two models for evaluations on street-level data and general-purpose image geolo
calization; the first model PIGEON is trained on data from the game of GeoGuessr
 and is capable of placing over 40% of its guesses within 25 kilometers of the t
arget location globally. We also develop a bot and deploy PIGEON in a blind expe
riment against humans ranking in the top 0.01% of players. We further challenge
one of the world's foremost professional GeoGuessr players to a series of six ma
tches with millions of viewers winning all six games. Our second model PIGEOTTO
differs in that it is trained on a dataset of images from Flickr and Wikipedia a
chieving state-of-the-art results on a wide range of image geolocalization bench
marks outperforming the previous SOTA by up to 7.7 percentage points on the city
 accuracy level and up to 38.8 percentage points on the country level. Our findi
ngs suggest that PIGEOTTO is the first image geolocalization model that effectiv
ely generalizes to unseen places and that our approach can pave the way for high
ly accurate planet-scale image geolocalization systems. Our code is available on
 GitHub.
****
Improving Spectral Snapshot Reconstruction with Spectral-Spatial Rectification
Jiancheng Zhang, Haijin Zeng, Yongyong Chen, Dengxiu Yu, Yin-Ping Zhao; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2024, pp. 25817-25826

How to effectively utilize the spectral and spatial characteristics of Hyperspec tral Image (HSI) is always a key problem in spectral snapshot reconstruction. Re cently the spectra-wise transformer has shown great potential in capturing inter -spectra similarities of HSI but the classic design of the transformer i.e. mult i-head division in the spectral (channel) dimension hinders the modeling of glob al spectral information and results in mean effect. In addition previous methods adopt the normal spatial priors without taking imaging processes into account a nd fail to address the unique spatial degradation in snapshot spectral reconstru ction. In this paper we analyze the influence of multi-head division and propose a novel Spectral-Spatial Rectification (SSR) method to enhance the utilization of spectral information and improve spatial degradation. Specifically SSR includ es two core parts: Window-based Spectra-wise Self-Attention (WSSA) and spAtial R ectification Block (ARB). WSSA is proposed to capture global spectral informatio n and account for local differences whereas ARB aims to mitigate the spatial deg radation using a spatial alignment strategy. The experimental results on simulat ion and real scenes demonstrate the effectiveness of the proposed modules and we also provide models at multiple scales to demonstrate the superiority of our ap proach.
****
Self-correcting LLM-controlled Diffusion Models
Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, Trevor Darrell; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6327-6336
Text-to-image generation has witnessed significant progress with the advent of d iffusion models. Despite the ability to generate photorealistic images current t ext-to-image diffusion models still often struggle to accurately interpret and f ollow complex input text prompts. In contrast to existing models that aim to gen erate images only with their best effort we introduce Self-correcting LLM-contro lled Diffusion (SLD). SLD is a framework that generates an image from the input prompt assesses its alignment with the prompt and performs self-corrections on t he inaccuracies in the generated image. Steered by an LLM controller SLD turns t ext-to-image generation into an iterative closed-loop process ensuring correctne ss in the resulting image. SLD is not only training-free but can also be seamles sly integrated with diffusion models behind API access such as DALL-E 3 to furth er boost the performance of state-of-the-art diffusion models. Experimental resu lts show that our approach can rectify a majority of incorrect generations parti cularly in generative numeracy attribute binding and spatial relationships. Furt hermore by simply adjusting the instructions to the LLM SLD can perform image ed iting tasks bridging the gap between text-to-image generation and image editing pipelines. Our code is available at: https://self-correcting-llm-diffusion.githu b.io.
****
PACER+: On-Demand Pedestrian Animation Controller in Driving Scenarios
Jingbo Wang, Zhengyi Luo, Ye Yuan, Yixuan Li, Bo Dai; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 718-72 8
We address the challenge of content diversity and controllability in pedestrian simulation for driving scenarios. Recent pedestrian animation frameworks have a significant limitation wherein they primarily focus on either following trajecto ry or the content of the reference video consequently overlooking the potential diversity of human motion within such scenarios. This limitation restricts the a bility to generate pedestrian behaviors that exhibit a wider range of variations and realistic motions and therefore restricts its usage to provide rich motion content for other components in the driving simulation system e.g. suddenly chan ged motion to which the autonomous vehicle should respond. In our approach we st rive to surpass the limitation by showcasing diverse human motions obtained from various sources such as generated human motions in addition to following the gi ven trajectory. The fundamental contribution of our framework lies in combining the motion tracking task with trajectory following which enables the tracking of specific motion parts (e.g. upper body) while simultaneously following the give

n trajectory by a single policy. This way we significantly enhance both the dive rsity of simulated human motion within the given scenario and the controllabilit y of the content including language-based control. Our framework facilitates the generation of a wide range of human motions contributing to greater realism and adaptability in pedestrian simulations for driving scenarios.
****
LTM: Lightweight Textured Mesh Extraction and Refinement of Large Unbounded Scen es for Efficient Storage and Real-time Rendering
Jaehoon Choi, Rajvi Shah, Qinbo Li, Yipeng Wang, Ayush Saraf, Changil Kim, Jia-B in Huang, Dinesh Manocha, Suhib Alsisan, Johannes Kopf; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5053 -5063
Advancements in neural signed distance fields (SDFs) have enabled modeling 3D su rface geometry from a set of 2D images of real-world scenes. Baking neural SDFs can extract explicit mesh with appearance baked into texture maps as neural feat ures. The baked meshes still have a large memory footprint and require a powerfu l GPU for real-time rendering. Neural optimization of such large meshes with dif ferentiable rendering pose significant challenges. We propose a method to produc e optimized meshes for large unbounded scenes with low triangle budget and high fidelity of geometry and appearance. We achieve this by combining advancements i n baking neural SDFs with classical mesh simplification techniques and proposing a joint appearance-geometry refinement step. The visual quality is comparable t o or better than state-of-the-art neural meshing and baking methods with high ge ometric accuracy despite significant reduction in triangle count making the prod uced meshes efficient for storage transmission and rendering on mobile hardware. We validate the effectiveness of the proposed method on large unbounded scenes from mip-NeRF 360 Tanks & Temples and Deep Blending datasets achieving at-par re ndering quality with 73x reduced triangles and 11x reduction in memory footprint .
****
Don't Drop Your Samples! Coherence-Aware Training Benefits Conditional Diffusion
Nicolas Dufour, Victor Besnier, Vicky Kalogeiton, David Picard; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6264-6273
Conditional diffusion models are powerful generative models that can leverage va rious types of conditional information such as class labels segmentation masks o r text captions. However in many real-world scenarios conditional information ma y be noisy or unreliable due to human annotation errors or weak alignment. In th is paper we propose the Coherence-Aware Diffusion (CAD) a novel method to integr ate confidence in conditional information into diffusion models allowing them to learn from noisy annotations without discarding data. We assume that each data point has an associated confidence score that reflects the quality of the condit ional information. We then condition the diffusion model on both the conditional information and the confidence score. In this way the model learns to ignore or discount the conditioning when the confidence is low. We show that our method i s theoretically sound and empirically effective on various conditional generatio n tasks. Moreover we show that leveraging confidence generates realistic and div erse samples that respect conditional information better than models trained on cleaned datasets where samples with low confidence have been discarded.
****
Flow-Guided Online Stereo Rectification for Wide Baseline Stereo
Anush Kumar, Fahim Mannan, Omid Hosseini Jafari, Shile Li, Felix Heide; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 15375-15385
Stereo rectification is widely considered "solved" due to the abundance of tradi tional approaches to perform rectification. However autonomous vehicles and robo ts in-the-wild require constant re-calibration due to exposure to various enviro nmental factors including vibration and structural stress when cameras are arran ged in a wide-baseline configuration. Conventional rectification methods fail in these challenging scenarios: especially for larger vehicles such as autonomous

freight trucks and semi-trucks the resulting incorrect rectification severely affects the quality of downstream tasks that use stereo/multi-view data. To tackle these challenges we propose an online rectification approach that operates at real-time rates while achieving high accuracy. We propose a novel learning-based online calibration approach that utilizes stereo correlation volumes built from a feature representation obtained from cross-image attention. Our model is trained to minimize vertical optical flow as proxy rectification constraint and predicts the relative rotation between the stereo pair. The method is real-time and even outperforms conventional methods used for offline calibration and substantially improves downstream stereo depth post-rectification. We release two public datasets (https://light.princeton.edu/online-stereo-recification/) a synthetic and experimental wide baseline dataset to foster further research.
****
DNGaussian: Optimizing Sparse-View 3D Gaussian Radiance Fields with Global-Local Depth Normalization
Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, Lin Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20775-20785
Radiance fields have demonstrated impressive performance in synthesizing novel views from sparse input views yet prevailing methods suffer from high training costs and slow inference speed. This paper introduces DNGaussian a depth-regularized framework based on 3D Gaussian radiance fields offering real-time and high-quality few-shot novel view synthesis at low costs. Our motivation stems from the highly efficient representation and surprising quality of the recent 3D Gaussian Splatting despite it will encounter a geometry degradation when input views decrease. In the Gaussian radiance fields we find this degradation in scene geometry primarily lined to the positioning of Gaussian primitives and can be mitigated by depth constraint. Consequently we propose a Hard and Soft Depth Regularization to restore accurate scene geometry under coarse monocular depth supervision while maintaining a fine-grained color appearance. To further refine detailed geometry reshaping we introduce Global-Local Depth Normalization enhancing the focus on small local depth changes. Extensive experiments on LLFF DTU and Blender datasets demonstrate that DNGaussian outperforms state-of-the-art methods achieving comparable or better results with significantly reduced memory cost a 25x reduction in training time and over 3000x faster rendering speed. Code is available at: https://github.com/Fictionarry/DNGaussian
****
ColorPCR: Color Point Cloud Registration with Multi-Stage Geometric-Color Fusion
Juncheng Mu, Lin Bie, Shaoyi Du, Yue Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21061-21070
Point cloud registration is still a challenging and open problem. For example when the overlap between two point clouds is extremely low geo-only features may be not sufficient. Therefore it is important to further explore how to utilize color data in this task. Under such circumstances we propose ColorPCR for color point cloud registration with multi-stage geometric-color fusion. We design a Hierarchical Color Enhanced Feature Extraction module to extract multi-level geometric-color features and a GeoColor Superpoint Matching Module to encode transformation-invariant geo-color global context for robust patch correspondences. In this way both geometric and color data can be used thus lead to robust performance even under extremely challenging scenarios such as low overlap between two point clouds. To evaluate the performance of our method we colorize 3DMatch/3DLoMatch datasets as Color3DMatch/Color3DLoMatch and evaluations on these datasets demonstrate the effectiveness of our proposed method. Our method achieves state-of-the-art registration recall of 97.5%/88.9% on them.
****
HomoFormer: Homogenized Transformer for Image Shadow Removal
Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25617-25626
The spatial non-uniformity and diverse patterns of shadow degradation conflict w

ith the weight sharing manner of dominant models which may lead to an unsatisfac
tory compromise. To tackle with this issue we present a novel strategy from the
view of shadow transformation in this paper: directly homogenizing the spatial d
istribution of shadow degradation. Our key design is the random shuffle operatio
n and its corresponding inverse operation. Specifically random shuffle operation
 stochastically rearranges the pixels across spatial space and the inverse opera
tion recovers the original order. After randomly shuffling the shadow diffuses i
n the whole image and the degradation appears in a homogenized way which can be
effectively processed by the local self-attention layer. Moreover we further dev
ise a new feed forward network with position modeling to exploit image structura
l information. Based on these elements we construct the final local window based
 transformer named HomoFormer for image shadow removal. Our HomoFormer can enjoy
 the linear complexity of local transformers while bypassing challenges of non-u
niformity and diversity of shadow. Extensive experiments are conducted to verify
 the superiority of our HomoFormer across public datasets.
****
What If the TV Was Off? Examining Counterfactual Reasoning Abilities of Multi-mo
dal Language Models
Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, Bingchen Zha
o; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2024, pp. 21853-21862
Counterfactual reasoning a fundamental aspect of human cognition involves contem
plating alternatives to established facts or past events significantly enhancing
 our abilities in planning and decision-making. In light of the advancements in
current multi-modal large language models we explore their effectiveness in coun
terfactual reasoning. To facilitate this investigation we introduce a novel data
set C-VQA specifically designed to examine the counterfactual reasoning capabili
ties of modern multi-modal large language models. This dataset is constructed by
 infusing original questions with counterfactual presuppositions spanning variou
s types such as numerical and boolean queries. It encompasses a mix of real and
synthetic data representing a wide range of difficulty levels. Our thorough eval
uations of contemporary vision-language models using this dataset have revealed
substantial performance drops with some models showing up to a 40% decrease high
lighting a significant gap between current models and human-like vision reasonin
g capabilities. We hope our dataset will serve as a vital benchmark for evaluati
ng the counterfactual reasoning capabilities of models. Code and dataset are pub
licly available at https://bzhao.me/C-VQA/.
****
What Do You See in Vehicle? Comprehensive Vision Solution for In-Vehicle Gaze Es
timation
Yihua Cheng, Yaning Zhu, Zongji Wang, Hongquan Hao, Yongwei Liu, Shiqing Cheng,
Xi Wang, Hyung Jin Chang; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2024, pp. 1556-1565
Driver's eye gaze holds a wealth of cognitive and intentional cues crucial for i
ntelligent vehicles. Despite its significance research on in-vehicle gaze estima
tion remains limited due to the scarcity of comprehensive and well-annotated dat
asets in real driving scenarios. In this paper we present three novel elements t
o advance in-vehicle gaze research. Firstly we introduce IVGaze a pioneering dat
aset capturing in-vehicle gaze collected from 125 individuals and covering a lar
ge range of gaze and head within vehicles. Conventional gaze collection systems
are inadequate for in-vehicle use. In this dataset we propose a new vision-based
 solution for in-vehicle gaze collection introducing a refined gaze target calib
ration method to tackle annotation challenges. Second our research focuses on in
-vehicle gaze estimation leveraging the IVGaze. Images of in-vehicle faces often
 suffer from low resolution prompting our introduction of a gaze pyramid transfo
rmer that harnesses transformer-based multilevel features integration. Expanding
 upon this we introduce the dual-stream gaze pyramid transformer (GazeDPTR). Emp
loying perspective transformation we rotate virtual cameras to normalize images
utilizing camera pose to merge normalized and original images for accurate gaze
estimation. GazeDPTR showcases state-of-the-art performance on the IVGaze datase

t. Thirdly we explore a novel strategy for gaze zone classification by extending the GazeDPTR. A foundational tri-plane and project gaze onto these planes are newly defined. Leveraging both positional features from the projection points and visual attributes from images we achieve superior performance compared to relying solely on visual features thereby substantiating the advantage of gaze estimation. The project is available at https://yihua.zone/work/ivgaze
****
Driving Everywhere with Large Language Model Policy Adaptation
Boyi Li, Yue Wang, Jiageng Mao, Boris Ivanovic, Sushant Veer, Karen Leung, Marco Pavone; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14948-14957
Adapting driving behavior to new environments customs and laws is a long-standing problem in autonomous driving precluding the widespread deployment of autonomous vehicles (AVs). In this paper we present LLaDA a simple yet powerful tool that enables human drivers and autonomous vehicles alike to drive everywhere by adapting their tasks and motion plans to traffic rules in new locations. LLaDA achieves this by leveraging the impressive zero-shot generalizability of large language models (LLMs) in interpreting the traffic rules in the local driver handbook. Through an extensive user study we show that LLaDA's instructions are useful in disambiguating in-the-wild unexpected situations. We also demonstrate LLaDA's ability to adapt AV motion planning policies in real-world datasets; LLaDA outperforms baseline planning approaches on all our metrics. Please check our website for more details: https://boyiliee.github.io/llada.
****
UFORecon: Generalizable Sparse-View Surface Reconstruction from Arbitrary and Unfavorable Sets
Youngju Na, Woo Jae Kim, Kyu Beom Han, Suhyeon Ha, Sung-Eui Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5094-5104
Generalizable neural implicit surface reconstruction aims to obtain an accurate underlying geometry given a limited number of multi-view images from unseen scenes. However existing methods select only informative and relevant views using predefined scores for training and testing phases. This constraint renders the model impractical in real-world scenarios where the availability of favorable combinations cannot always be ensured. We introduce and validate a view-combination score to indicate the effectiveness of the input view combination. We observe that previous methods output degenerate solutions under arbitrary and unfavorable sets. Building upon this finding we propose UFORecon a robust view-combination generalizable surface reconstruction framework. To achieve this we apply cross-view matching transformers to model interactions between source images and build correlation frustums to capture global correlations. Additionally we explicitly encode pairwise feature similarities as view-consistent priors. Our proposed framework significantly outperforms previous methods in terms of view-combination generalizability and also in the conventional generalizable protocol trained with favorable view-combinations. The code is available at https://github.com/Youngju-Na/UFORecon.
****
FAR: Flexible Accurate and Robust 6DoF Relative Camera Pose Estimation
Chris Rockwell, Nilesh Kulkarni, Linyi Jin, Jeong Joon Park, Justin Johnson, David F. Fouhey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19854-19864
Estimating relative camera poses between images has been a central problem in computer vision. Methods that find correspondences and solve for the fundamental matrix offer high precision in most cases. Conversely methods predicting pose directly using neural networks are more robust to limited overlap and can infer absolute translation scale but at the expense of reduced precision. We show how to combine the best of both methods; our approach yields results that are both precise and robust while also accurately inferring translation scales. At the heart of our model lies a Transformer that (1) learns to balance between solved and learned pose estimations and (2) provides a prior to guide a solver. A comprehensi

ve analysis supports our design choices and demonstrates that our method adapts flexibly to various feature extractors and correspondence estimators showing state-of-the-art performance in 6DoF pose estimation on Matterport3D InteriorNet StreetLearn and Map-free Relocalization.
****
eTraM: Event-based Traffic Monitoring Dataset
Aayush Atul Verma, Bharatesh Chakravarthi, Arpitsinh Vaghela, Hua Wei, Yezhou Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22637-22646
Event cameras with their high temporal and dynamic range and minimal memory usage have found applications in various fields. However their potential in static traffic monitoring remains largely unexplored. To facilitate this exploration we present eTraM - a first-of-its-kind fully event-based traffic monitoring dataset. eTraM offers 10 hr of data from different traffic scenarios in various lighting and weather conditions providing a comprehensive overview of real-world situations. Providing 2M bounding box annotations it covers eight distinct classes of traffic participants ranging from vehicles to pedestrians and micro-mobility. eTraM's utility has been assessed using state-of-the-art methods for traffic participant detection including RVT RED and YOLOv8. We quantitatively evaluate the ability of event-based models to generalize on nighttime and unseen scenes. Our findings substantiate the compelling potential of leveraging event cameras for traffic monitoring opening new avenues for research and application. eTraM is available at https://eventbasedvision.github.io/eTraM.
****
MoCha-Stereo: Motif Channel Attention Network for Stereo Matching
Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, Jia Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27768-27777
Learning-based stereo matching techniques have made significant progress. However existing methods inevitably lose geometrical structure information during the feature channel generation process resulting in edge detail mismatches. In this paper the Motif Channel Attention Stereo Matching Network (MoCha-Stereo) is designed to address this problem. We provide the Motif Channel Correlation Volume (MCCV) to determine more accurate edge matching costs. MCCV is achieved by projecting motif channels which capture common geometric structures in feature channels onto feature maps and cost volumes. In addition edge variations in the reconstruction error map also affect details matching we propose the Reconstruction Error Motif Penalty (REMP) module to further refine the full-resolution disparity estimation. REMP integrates the frequency information of typical channel features from the reconstruction error. MoCha-Stereo ranks 1st on the KITTI-2015 and KITTI-2012 Reflective leaderboards. Our structure also shows excellent performance in Multi-View Stereo. Code is avaliable at https://github.com/ZYangChen/MoCha-Stereo.
****
Koala: Key Frame-Conditioned Long Video-LLM
Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A. Plummer, Bryan Russell, Kate Saenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13581-13591
Long video question answering is a challenging task that involves recognizing short-term activities and reasoning about their fine-grained relationships. State-of-the-art video Large Language Models (vLLMs) hold promise as a viable solution due to their demonstrated emergent capabilities on new tasks. However despite being trained on millions of short seconds-long videos vLLMs are unable to understand minutes-long videos and accurately answer questions about them. To address this limitation we propose a lightweight and self-supervised approach Key frame-conditioned long video-LLM (Koala) that introduces learnable spatiotemporal queries to adapt pretrained vLLMs for generalizing to longer videos. Our approach introduces two new tokenizers that condition on visual tokens computed from sparse video key frames for understanding short and long video moments. We train our proposed approach on HowTo100M and demonstrate its effectiveness on zero-shot lon

g video understanding benchmarks where it outperforms state-of-the-art large mod els by 3 - 6% in absolute accuracy across all tasks. Surprisingly we also empiri cally show that our approach not only helps a pretrained vLLM to understand long videos but also improves its accuracy on short-term action recognition.
****
Extend Your Own Correspondences: Unsupervised Distant Point Cloud Registration b y Progressive Distance Extension
Quan Liu, Hongzi Zhu, Zhenxi Wang, Yunsong Zhou, Shan Chang, Minyi Guo; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 20816-20826
Registration of point clouds collected from a pair of distant vehicles provides a comprehensive and accurate 3D view of the driving scenario which is vital for driving safety related applications yet existing literature suffers from the exp ensive pose label acquisition and the deficiency to generalize to new data distr ibutions. In this paper we propose EYOC an unsupervised distant point cloud regi stration method that adapts to new point cloud distributions on the fly requirin g no global pose labels. The core idea of EYOC is to train a feature extractor i n a progressive fashion where in each round the feature extractor trained with n ear point cloud pairs can label slightly farther point cloud pairs enabling self -supervision on such far point cloud pairs. This process continues until the der ived extractor can be used to register distant point clouds. Particularly to ena ble high-fidelity correspondence label generation we devise an effective spatial filtering scheme to select the most representative correspondences to register a point cloud pair and then utilize the aligned point clouds to discover more co rrect correspondences. Experiments show that EYOC can achieve comparable perform ance with state-of-the-art supervised methods at a lower training cost. Moreover it outwits supervised methods regarding generalization performance on new data distributions.
****
HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucinatio n and Visual Illusion in Large Vision-Language Models
Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun W ang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, Tianyi Zhou; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2024, pp. 14375-14385
We introduce "HallusionBench" a comprehensive benchmark designed for the evaluat ion of image-context reasoning. This benchmark presents significant challenges t o advanced large visual-language models (LVLMs) such as GPT-4V(ision) Gemini Pro Vision Claude 3 and LLaVA-1.5 by emphasizing nuanced understanding and interpre tation of visual data. The benchmark comprises 346 images paired with 1129 quest ions all meticulously crafted by human experts. We introduce a novel structure f or these visual questions designed to establish control groups. This structure e nables us to conduct a quantitative analysis of the models' response tendencies logical consistency and various failure modes. In our evaluation on HallusionBen ch we benchmarked 15 different models highlighting a 31.42% question-pair accura cy achieved by the state-of-the-art GPT-4V. Notably all other evaluated models a chieve accuracy below 16%. Moreover our analysis not only highlights the observe d failure modes including language hallucination and visual illusion but also de epens an under standing of these pitfalls. Our comprehensive case studies within HallusionBench shed light on the challenges of hallucination and illusion in LV LMs. Based on these insights we suggest potential pathways for their future impr ovement. The benchmark and codebase can be accessed at https://github.com/tianyi lab/HallusionBench.
****
ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection
Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, Changqing Zhang; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17480-17489
Out-of-distribution (OOD) detection methods often exploit auxiliary outliers to train model identifying OOD samples especially discovering challenging outliers

from auxiliary outliers dataset to improve OOD detection. However they may still face limitations in effectively distinguishing between the most challenging OOD samples that are much like in-distribution (ID) data i.e. ID-like samples. To this end we propose a novel OOD detection framework that discovers ID-like outliers using CLIP from the vicinity space of the ID samples thus helping to identify these most challenging OOD samples. Then a prompt learning framework is proposed that utilizes the identified ID-like outliers to further leverage the capabilities of CLIP for OOD detection. Benefiting from the powerful CLIP we only need a small number of ID samples to learn the prompts of the model without exposing other auxiliary outlier datasets. By focusing on the most challenging ID-like OOD samples and elegantly exploiting the capabilities of CLIP our method achieves superior few-shot learning performance on various real-world image datasets (e.g. in 4-shot OOD detection on the ImageNet-1k dataset our method reduces the average FPR95 by 12.16% and improves the average AUROC by 2.76% compared to state-of-the-art methods).
****
Breathing Life Into Sketches Using Text-to-Video Priors
Rinon Gal, Yael Vinker, Yuval Alaluf, Amit Bermano, Daniel Cohen-Or, Ariel Shamir, Gal Chechik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4325-4336
A sketch is one of the most intuitive and versatile tools humans use to convey their ideas visually. An animated sketch opens another dimension to the expression of ideas and is widely used by designers for a variety of purposes. Animating sketches is a laborious process requiring extensive experience and professional design skills. In this work we present a method that automatically adds motion to a single-subject sketch (hence "breathing life into it") merely by providing a text prompt indicating the desired motion. The output is a short animation provided in vector representation which can be easily edited. Our method does not require extensive training but instead leverages the motion prior of a large pretrained text-to-video diffusion model using a score-distillation loss to guide the placement of strokes. To promote natural and smooth motion and to better preserve the sketch's appearance we model the learned motion through two components. The first governs small local deformations and the second controls global affine transformations. Surprisingly we find that even models that struggle to generate sketch videos on their own can still serve as a useful backbone for animating abstract representations.
****
Multi-modal Learning for Geospatial Vegetation Forecasting
Vitus Benson, Claire Robin, Christian Requena-Mesa, Lazaro Alonso, Nuno Carvalhais, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, Markus Reichstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27788-27799
Precise geospatial vegetation forecasting holds potential across diverse sectors including agriculture forestry humanitarian aid and carbon accounting. To leverage the vast availability of satellite imagery for this task various works have applied deep neural networks for predicting multispectral images in photorealistic quality. However the important area of vegetation dynamics has not been thoroughly explored. Our study introduces GreenEarthNet the first dataset specifically designed for high-resolution vegetation forecasting and Contextformer a novel deep learning approach for predicting vegetation greenness from Sentinel 2 satellite images with fine resolution across Europe. Our multi-modal transformer model Contextformer leverages spatial context through a vision backbone and predicts the temporal dynamics on local context patches incorporating meteorological time series in a parameter-efficient manner. The GreenEarthNet dataset features a learned cloud mask and an appropriate evaluation scheme for vegetation modeling. It also maintains compatibility with the existing satellite imagery forecasting dataset EarthNet2021 enabling cross-dataset model comparisons. Our extensive qualitative and quantitative analyses reveal that our methods outperform a broad range of baseline techniques. This includes surpassing previous state-of-the-art models on EarthNet2021 as well as adapted models from time series forecasting and

video prediction. To the best of our knowledge this work presents the first mod
els for continental-scale vegetation modeling at fine resolution able to capture
 anomalies beyond the seasonal cycle thereby paving the way for predicting veget
ation health and behaviour in response to climate variability and extremes. We p
rovide open source code and pre-trained weights to reproduce our experimental re
sults under https://github.com/vitusbenson/greenearthnet.
****

Learning Diffusion Texture Priors for Image Restoration
Tian Ye, Sixiang Chen, Wenhao Chai, Zhaohu Xing, Jing Qin, Ge Lin, Lei Zhu; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 2524-2534
Diffusion Models have shown remarkable performance in image generation tasks whi
ch are capable of generating diverse and realistic image content. When adopting
diffusion models for image restoration the crucial challenge lies in how to pres
erve high-level image fidelity in the randomness diffusion process and generate
accurate background structures and realistic texture details. In this paper we p
ropose a general framework and develop a Diffusion Texture Prior Model (DTPM) fo
r image restoration tasks. DTPM explicitly models high-quality texture details t
hrough the diffusion process rather than global contextual content. In phase one
 of the training stage we pre-train DTPM on approximately 55K high-quality image
 samples after which we freeze most of its parameters. In phase two we insert co
nditional guidance adapters into DTPM and equip it with an initial predictor the
reby facilitating its rapid adaptation to downstream image restoration tasks. Ou
r DTPM could mitigate the randomness of traditional diffusion models by utilizin
g encapsulated rich and diverse texture knowledge and background structural info
rmation provided by the initial predictor during the sampling process. Our compr
ehensive evaluations of five image restoration tasks demonstrate DTPM's superior
ity over existing regression and diffusion-based image restoration methods in pe
rceptual quality and its exceptional generalization capabilities.
****

Bring Event into RGB and LiDAR: Hierarchical Visual-Motion Fusion for Scene Flow
Hanyu Zhou, Yi Chang, Zhiwei Shi; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2024, pp. 26477-26486
Single RGB or LiDAR is the mainstream sensor for the challenging scene flow whic
h relies heavily on visual features to match motion features. Compared with sing
le modality existing methods adopt a fusion strategy to directly fuse the cross-
modal complementary knowledge in motion space. However these direct fusion metho
ds may suffer the modality gap due to the visual intrinsic heterogeneous nature
between RGB and LiDAR thus deteriorating motion features. We discover that event
 has the homogeneous nature with RGB and LiDAR in both visual and motion spaces.
 In this work we bring the event as a bridge between RGB and LiDAR and propose a
 novel hierarchical visual-motion fusion framework for scene flow which explores
 a homogeneous space to fuse the cross-modal complementary knowledge for physica
l interpretation. In visual fusion we discover that event has a complementarity
(relative v.s. absolute) in luminance space with RGB for high dynamic imaging an
d has a complementarity (local boundary v.s. global shape) in scene structure sp
ace with LiDAR for structure integrity. In motion fusion we figure out that RGB
event and LiDAR are complementary (spatial-dense temporal-dense v.s. spatiotempo
ral-sparse) to each other in correlation space which motivates us to fuse their
motion correlations for motion continuity. The proposed hierarchical fusion can
explicitly fuse the multimodal knowledge to progressively improve scene flow fro
m visual space to motion space. Extensive experiments have been performed to ver
ify the superiority of the proposed method.
****

Entangled View-Epipolar Information Aggregation for Generalizable Neural Radianc
e Fields
Zhiyuan Min, Yawei Luo, Wei Yang, Yuesong Wang, Yi Yang; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 490
6-4916
Generalizable NeRF can directly synthesize novel views across new scenes elimina

ting the need for scene-specific retraining in vanilla NeRF. A critical enabling factor in these approaches is the extraction of a generalizable 3D representation by aggregating source-view features. In this paper we propose an Entangled View-Epipolar Information Aggregation method dubbed EVE-NeRF. Different from existing methods that consider cross-view and along-epipolar information independently EVE-NeRF conducts the view-epipolar feature aggregation in an entangled manner by injecting the scene-invariant appearance continuity and geometry consistency priors to the aggregation process. Our approach effectively mitigates the potential lack of inherent geometric and appearance constraint resulting from one-dimensional interactions thus further boosting the 3D representation generalizablity. EVE-NeRF attains state-of-the-art performance across various evaluation scenarios. Extensive experiments demonstate that compared to prevailing single-dimensional aggregation the entangled network excels in the accuracy of 3D scene geometry and appearance reconstruction. Our code is publicly available at https://github.com/tatakai1/EVENeRF.
****
Jack of All Tasks Master of Many: Designing General-Purpose Coarse-to-Fine Vision-Language Model
Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, Amjad Almahairi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14076-14088
The ability of large language models (LLMs) to process visual inputs has given rise to general-purpose vision systems unifying various vision-language (VL) tasks by instruction tuning. However due to the enormous diversity in input-output formats in the vision domain existing general-purpose models fail to successfully integrate segmentation and multi-image inputs with coarse-level tasks into a single framework. In this work we introduce VistaLLM a powerful visual system that addresses coarse- and fine grained VL tasks over single and multiple input images using a unified framework. VistaLLM utilizes an instruction-guided image tokenizer that filters global embeddings using task descriptions to extract compressed and refined features from numerous images. Moreover VistaLLM employs a gradient-aware adaptive sampling technique to represent binary segmentation masks as sequences significantly improving over previously used uniform sampling. To bolster the desired capability of VistaLLM we curate CoinIt a comprehensive coarse-to-fine instruction tuning dataset with 6.8M samples. We also address the lack of multi-image grounding datasets by introducing a novel task AttCoSeg (Attribute-level Co Segmentation) which boosts the model's reasoning and grounding capability over multiple input images. Extensive experiments on a wide range of V- and VL tasks demonstrate the effectiveness of VistaLLM by achieving consistent state-of-the-art performance over strong baselines across many downstream tasks. Our project page can be found at https://shramanpramanick.github.io/VistaLLM/
****
MMVP: A Multimodal MoCap Dataset with Vision and Pressure Sensors
He Zhang, Shenghao Ren, Haolei Yuan, Jianhui Zhao, Fan Li, Shuangpeng Sun, Zhenghao Liang, Tao Yu, Qiu Shen, Xun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21842-21852
Foot contact is an important cue for human motion capture understanding and generation. Existing datasets tend to annotate dense foot contact using visual matching with thresholding or incorporating pressure signals. However these approaches either suffer from low accuracy or are only designed for small-range and slow motion. There is still a lack of a vision-pressure multimodal dataset with large-range and fast human motion as well as accurate and dense foot-contact annotation. To fill this gap we propose a Multimodal MoCap Dataset with Vision and Pressure sensors named MMVP. MMVP provides accurate and dense plantar pressure signals synchronized with RGBD observations which is especially useful for both plausible shape estimation robust pose fitting without foot drifting and accurate global translation tracking. To validate the dataset we propose an RGBD-P SMPL fitting method and also a monocular-video-based baseline framework VP-MoCap for human motion capture. Experiments demonstrate that our RGBD-P SMPL Fitting results significantly outperform pure visual motion capture. Moreover VP-MoCap outperforms

SOTA methods in foot-contact and global translation estimation accuracy. We bel ieve the configuration of the dataset and the baseline frameworks will stimulate the research in this direction and also provide a good reference for MoCap appl ications in various domains. Project page: https://metaverse-ai-lab-thu.github.i o/MMVP-Dataset/.
****
YolOOD: Utilizing Object Detection Concepts for Multi-Label Out-of-Distribution Detection
Alon Zolfi, Guy Amit, Amit Baras, Satoru Koda, Ikuya Morikawa, Yuval Elovici, As af Shabtai; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2024, pp. 5788-5797
Out-of-distribution (OOD) detection has attracted a large amount of attention fr om the machine learning research community in recent years due to its importance in deployed systems. Most of the previous studies focused on the detection of O OD samples in the multi-class classification task. However OOD detection in the multi-label classification task a more common real-world use case remains an und erexplored domain. In this research we propose YolOOD - a method that utilizes c oncepts from the object detection domain to perform OOD detection in the multi-l abel classification task. Object detection models have an inherent ability to di stinguish between objects of interest (in-distribution data) and irrelevant obje cts (OOD data) in images that contain multiple objects belonging to different cl ass categories. These abilities allow us to convert a regular object detection m odel into an image classifier with inherent OOD detection capabilities with just minor changes. We compare our approach to state-of-the-art OOD detection method s and demonstrate YolOOD's ability to outperform these methods on a comprehensiv e suite of in-distribution and OOD benchmark datasets.
****
SchurVINS: Schur Complement-Based Lightweight Visual Inertial Navigation System
Yunfei Fan, Tianyu Zhao, Guidong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17964-17973
Accuracy and computational efficiency are the most important metrics to Visual I nertial Navigation System (VINS). The existing VINS algorithms with either high accuracy or low computational complexity are difficult to provide the high preci sion localization in resource-constrained devices. To this end we propose a nove l filter-based VINS framework named SchurVINS (SV) which could guarantee both hi gh accuracy by building a complete residual model and low computational complexi ty with Schur complement. Technically we first formulate the full residual model where Gradient Hessian and observation covariance are explicitly modeled. Then Schur complement is employed to decompose the full model into ego-motion residua l model and landmark residual model. Finally Extended Kalman Filter (EKF) update is implemented in these two models with high efficiency. Experiments on EuRoC a nd TUM-VI datasets show that our method notably outperforms state-of-the-art (SO TA) methods in both accuracy and computational complexity. The experimental code of SchurVINS is available at https://github.com/bytedance/SchurVINS.
****
Collaborating Foundation Models for Domain Generalized Semantic Segmentation
Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, Stéphane Lathuiliè re; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2024, pp. 3108-3119
Domain Generalized Semantic Segmentation (DGSS) deals with training a model on a labeled source domain with the aim of generalizing to unseen domains during inf erence. Existing DGSS methods typically effectuate robust features by means of D omain Randomization (DR). Such an approach is often limited as it can only accou nt for style diversification and not content. In this work we take an orthogonal approach to DGSS and propose to use an assembly of CoLlaborative FOUndation mod els for Domain Generalized Semantic Segmentation (CLOUDS). In detail CLOUDS is a framework that integrates Foundation Models of various kinds: (i) CLIP backbone for its robust feature representation (ii) Diffusion Model to diversify the con tent thereby covering various modes of the possible target distribution and (iii) Segment Anything Model (SAM) for iteratively refining the predictions of the s

egmentation model. Extensive experiments show that our CLOUDS excels in adapting from synthetic to real DGSS benchmarks and under varying weather conditions notably outperforming prior methods by 5.6% and 6.7% on averaged mIoU respectively. Our code is available at https://github.com/yasserben/CLOUDS
****
Towards Variable and Coordinated Holistic Co-Speech Motion Generation
Yifei Liu, Qiong Cao, Yandong Wen, Huaiguang Jiang, Changxing Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1566-1576
This paper addresses the problem of generating lifelike holistic co-speech motions for 3D avatars focusing on two key aspects: variability and coordination. Variability allows the avatar to exhibit a wide range of motions even with similar speech content while coordination ensures a harmonious alignment among facial expressions hand gestures and body poses. We aim to achieve both with ProbTalk a unified probabilistic framework designed to jointly model facial hand and body movements in speech. ProbTalk builds on the variational autoencoder (VAE) architecture and incorporates three core designs. First we introduce product quantization (PQ) to the VAE which enriches the representation of complex holistic motion. Second we devise a novel non-autoregressive model that embeds 2D positional encoding into the product-quantized representation thereby preserving essential structure information of the PQ codes. Last we employ a secondary stage to refine the preliminary prediction further sharpening the high-frequency details. Coupling these three designs enables ProbTalk to generate natural and diverse holistic co-speech motions outperforming several state-of-the-art methods in qualitative and quantitative evaluations particularly in terms of realism. Our code and model will be released for research purposes at https://feifeifeiliu.github.io/probtalk/.
****
JoAPR: Cleaning the Lens of Prompt Learning for Vision-Language Models
Yuncheng Guo, Xiaodong Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28695-28705
Leveraging few-shot datasets in prompt learning for Vision-Language Models eliminates the need for manual prompt engineering while highlighting the necessity of accurate annotations for the labels. However high-level or complex label noise challenges prompt learning for Vision-Language Models. Aiming at this issue we propose a new framework for improving its robustness. Specifically we introduce the Joint Adaptive Partitioning for Label Refurbishment (JoAPR) a structured framework encompassing two key steps. 1) Data Partitioning where we differentiate between clean and noisy data using joint adaptive thresholds. 2) Label Refurbishment where we correct the labels based on the partition outcomes before retraining the network. Our comprehensive experiments confirm that JoAPR substantially enhances the robustness of prompt learning for Vision-Language Models against label noise offering a promising direction for future research.
****
AllSpark: Reborn Labeled Features from Unlabeled in Transformer for Semi-Supervised Semantic Segmentation
Haonan Wang, Qixiang Zhang, Yi Li, Xiaomeng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3627-3636
Semi-supervised semantic segmentation (SSSS) has been proposed to alleviate the burden of time-consuming pixel-level manual labeling which leverages limited labeled data along with larger amounts of unlabeled data. Current state-of-the-art methods train the labeled data with ground truths and unlabeled data with pseudo labels. However the two training flows are separate which allows labeled data to dominate the training process resulting in low-quality pseudo labels and consequently sub-optimal results. To alleviate this issue we present AllSpark which reborns the labeled features from unlabeled ones with the channel-wise cross-attention mechanism. We further introduce a Semantic Memory along with a Channel Semantic Grouping strategy to ensure that unlabeled features adequately represent labeled features. The AllSpark shed new light on the architecture level designs of SSSS rather than framework level which avoids increasingly complicated trainin

g pipeline designs. It can also be regarded as a flexible bottleneck module that
 can be seamlessly integrated into a general transformer-based segmentation mode
l. The proposed AllSpark outperforms existing methods across all evaluation prot
ocols on Pascal Cityscapes and COCO benchmarks without bells-and-whistles. Code
and model weights are available at: https://github.com/xmed-lab/AllSpark.
****
Open-Vocabulary 3D Semantic Segmentation with Foundation Models
Li Jiang, Shaoshuai Shi, Bernt Schiele; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21284-21294
In dynamic 3D environments the ability to recognize a diverse range of objects w
ithout the constraints of predefined categories is indispensable for real-world
applications. In response to this need we introduce OV3D an innovative framework
 designed for open-vocabulary 3D semantic segmentation. OV3D leverages the broad
 open-world knowledge embedded in vision and language foundation models to estab
lish a fine-grained correspondence between 3D points and textual entity descript
ions. These entity descriptions are enriched with contextual information enablin
g a more open and comprehensive understanding. By seamlessly aligning 3D point f
eatures with entity text features OV3D empowers open-vocabulary recognition in t
he 3D domain achieving state-of-the-art open-vocabulary semantic segmentation pe
rformance across multiple datasets including ScanNet Matterport3D and nuScenes.
****
SIGNeRF: Scene Integrated Generation for Neural Radiance Fields
Jan-Niklas Dihlmann, Andreas Engelhardt, Hendrik Lensch; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 667
9-6688
Advances in image diffusion models have recently led to notable improvements in
the generation of high-quality images. In combination with Neural Radiance Field
s (NeRFs) they enabled new opportunities in 3D generation. However most generati
ve 3D approaches are object-centric and applying them to editing existing photor
ealistic scenes is not trivial. We propose SIGNeRF a novel approach for fast and
 controllable NeRF scene editing and scene-integrated object generation. A new g
enerative update strategy ensures 3D consistency across the edited images withou
t requiring iterative optimization. We find that depth-conditioned diffusion mod
els inherently possess the capability to generate 3D consistent views by request
ing a grid of images instead of single views. Based on these insights we introdu
ce a multi-view reference sheet of modified images. Our method updates an image
collection consistently based on the reference sheet and refines the original Ne
RF with the newly generated image set in one go. By exploiting the depth conditi
oning mechanism of the image diffusion model we gain fine control over the spati
al location of the edit and enforce shape guidance by a selected region or an ex
ternal mesh.
****
ViP-LLaVA: Making Large Multimodal Models Understand Arbitrary Visual Prompts
Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, De
nnis Park, Yong Jae Lee; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 12914-12923
While existing large vision-language multimodal models focus on whole image unde
rstanding there is a prominent gap in achieving region-specific comprehension. C
urrent approaches that use textual coordinates or spatial encodings often fail t
o provide a user-friendly interface for visual prompting. To address this challe
nge we introduce a novel multimodal model capable of decoding arbitrary (free-fo
rm) visual prompts. This allows users to intuitively mark images and interact wi
th the model using natural cues like a "red bounding box" or "pointed arrow'". O
ur simple design directly overlays visual markers onto the RGB image eliminating
 the need for complex region encodings yet achieves state-of-the-art performance
 on region-understanding tasks like Visual7W PointQA and Visual Commonsense Reas
oning benchmark. Furthermore we present ViP-Bench a comprehensive benchmark to a
ssess the capability of models in understanding visual prompts across multiple d
imensions enabling future research in this domain. Code data and model are publi
cly available.

****
OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabular
y Detection and StructurEd Representation
Ganlong Zhao, Guanbin Li, Weikai Chen, Yizhou Yu; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16296-1630
6
Recent advances in Iterative Vision-and-Language Navigation(IVLN) introduce a mo
re meaningful and practical paradigm of VLN by maintaining the agent's memory ac
ross tours of scenes. Although the long-term memory aligns better with the persi
stent nature of the VLN task it poses more challenges on how to utilize the high
ly unstructured navigation memory with extremely sparse supervision. Towards thi
s end we propose OVER-NAV which aims to go over and beyond the current arts of I
VLN techniques. In particular we propose to incorporate LLMs and open-vocabulary
 detectors to distill key information and establish correspondence between multi
-modal signals. Such a mechanism introduces reliable cross-modal supervision and
 enables on-the-fly generalization to unseen scenes without the need of extra an
notation and re-training. To fully exploit the interpreted navigation data we fu
rther introduce a structured representation coded Omnigraph to effectively integ
rate multi-modal information along the tour. Accompanied with a novel omnigraph
fusion mechanism OVER-NAV is able to extract the most relevant knowledge from om
nigraph for a more accurate navigating action. In addition OVER-NAV seamlessly s
upports both discrete and continuous environments under a unified framework. We
demonstrate the superiority of OVER-NAV in extensive experiments.
****
1-Lipschitz Layers Compared: Memory Speed and Certifiable Robustness
Bernd Prach, Fabio Brau, Giorgio Buttazzo, Christoph H. Lampert; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 24574-24583
The robustness of neural networks against input perturbations with bounded magni
tude represents a serious concern in the deployment of deep learning models in s
afety-critical systems. Recently the scientific community has focused on enhanci
ng certifiable robustness guarantees by crafting \ols neural networks that lever
age Lipschitz bounded dense and convolutional layers. Different methods have bee
n proposed in the literature to achieve this goal however comparing the performa
nce of such methods is not straightforward since different metrics can be releva
nt (e.g. training time memory usage accuracy certifiable robustness) for differe
nt applications. Therefore this work provides a thorough comparison between diff
erent methods covering theoretical aspects such as computational complexity and
memory requirements as well as empirical measurements of time per epoch required
 memory accuracy and certifiable robust accuracy. The paper also provides some g
uidelines and recommendations to support the user in selecting the methods that
work best depending on the available resources. We provide code at github.com/be
rndprach/1LipschitzLayersCompared
****
All Rivers Run to the Sea: Private Learning with Asymmetric Flows
Yue Niu, Ramy E. Ali, Saurav Prakash, Salman Avestimehr; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 123
53-12362
Data privacy is of great concern in cloud machine-learning service platforms whe
n sensitive data are exposed to service providers. While private computing envir
onments (e.g. secure enclaves) and cryptographic approaches (e.g. homomorphic en
cryption) provide strong privacy protection their computing performance still fa
lls short compared to cloud GPUs. To achieve privacy protection with high comput
ing performance we propose Delta a new private training and inference framework
with comparable model performance as non-private centralized training. Delta fea
tures two asymmetric data flows: the main information-sensitive flow and the res
idual flow. The main part flows into a small model while the residuals are offlo
aded to a large model. Specifically Delta embeds the information-sensitive repre
sentations into a low-dimensional space while pushing the information-insensitiv
e part into high-dimension residuals. To ensure privacy protection the low-dimen

sional information-sensitive part is secured and fed to a small model in a priva
te environment. On the other hand the residual part is sent to fast cloud GPUs a
nd processed by a large model. To further enhance privacy and reduce the communi
cation cost Delta applies a random binary quantization technique along with a DP
-based technique to the residuals before sharing them with the public platform.
We theoretically show that Delta guarantees differential privacy in the public e
nvironment and greatly reduces the complexity in the private environment. We con
duct empirical analyses on CIFAR-10 CIFAR-100 and ImageNet datasets and ResNet-1
8 and ResNet-34 showing that Delta achieves strong privacy protection fast train
ing and inference without significantly compromising the model utility.
****
Generating Illustrated Instructions
Sachit Menon, Ishan Misra, Rohit Girdhar; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6274-6284
We introduce a new task of generating "Illustrated Instructions" i.e. visual ins
tructions customized to a user's needs. We identify desiderata unique to this ta
sk and formalize it through a suite of automatic and human evaluation metrics de
signed to measure the validity consistency and efficacy of the generations. We c
ombine the power of large language models (LLMs) together with strong text-to-im
age generation diffusion models to propose a simple approach called StackedDiffu
sion which generates such illustrated instructions given text as input. The resu
lting model strongly outperforms baseline approaches and state-of-the-art multim
odal LLMs; and in 30% of cases users even prefer it to human-generated articles.
 Most notably it enables various new and exciting applications far beyond what s
tatic articles on the web can provide such as personalized instructions complete
 with intermediate steps and pictures in response to a user's individual situati
on.
****
Construct to Associate: Cooperative Context Learning for Domain Adaptive Point C
loud Segmentation
Guangrui Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2024, pp. 27917-27926
This paper tackles the domain adaptation problem in point cloud semantic segment
ation which performs adaptation from a fully labeled domain (source domain) to a
n unlabeled target domain. Due to the unordered property of point clouds LiDAR s
cans typically show varying geometric structures across different regions in ter
ms of density noises etc hence leading to increased dynamics on context. However
 such characteristics are not consistent across domains due to the difference in
 sensors environments etc thus hampering the effective scene comprehension acros
s domains. To solve this we propose Cooperative Context Learning that performs c
ontext modeling and modulation from different aspects but in a cooperative manne
r. Specifically we first devise context embeddings to discover and model context
ual relationships with close neighbors in a learnable manner. Then with the cont
ext embeddings from two domains we introduce a set of learnable prototypes to at
tend and associate them under the attention paradigm. As a result these prototyp
es naturally establish long-range dependency across regions and domains thereby
encouraging the transfer of context knowledge and easing the adaptation. Moreove
r the attention in turn attunes and guides the local context modeling and urges
them to focus on the domain-invariant context knowledge thus promoting the adapt
ation in a cooperative manner. Experiments on representative benchmarks verify t
hat our method attains the new state-of-the-art.
****
Robust Image Denoising through Adversarial Frequency Mixup
Donghun Ryou, Inju Ha, Hyewon Yoo, Dongwan Kim, Bohyung Han; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 2723-2732
Image denoising approaches based on deep neural networks often struggle with ove
rfitting to specific noise distributions present in training data. This challeng
e persists in existing real-world denoising networks which are trained using a l
imited spectrum of real noise distributions and thus show poor robustness to out

-of-distribution real noise types. To alleviate this issue we develop a novel tr aining framework called Adversarial Frequency Mixup (AFM). AFM leverages mixup i n the frequency domain to generate noisy images with distinctive and challenging noise characteristics all the while preserving the properties of authentic real -world noise. Subsequently incorporating these noisy images into the training pi peline enhances the denoising network's robustness to variations in noise distri butions. Extensive experiments and analyses conducted on a wide range of real no ise benchmarks demonstrate that denoising networks trained with our proposed fra mework exhibit significant improvements in robustness to unseen noise distributi ons. The code is available at https://github.com/dhryougit/AFM.
****
HandBooster: Boosting 3D Hand-Mesh Reconstruction by Conditional Synthesis and S ampling of Hand-Object Interactions
Hao Xu, Haipeng Li, Yinqiao Wang, Shuaicheng Liu, Chi-Wing Fu; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p p. 10159-10169
Reconstructing 3D hand mesh robustly from a single image is very challenging due to the lack of diversity in existing real-world datasets. While data synthesis helps relieve the issue the syn-to-real gap still hinders its usage. In this wor k we present HandBooster a new approach to uplift the data diversity and boost t he 3D hand-mesh reconstruction performance by training a conditional generative space on hand-object interactions and purposely sampling the space to synthesize effective data samples. First we construct versatile content-aware conditions t o guide a diffusion model to produce realistic images with diverse hand appearan ces poses views and backgrounds; favorably accurate 3D annotations are obtained for free. Then we design a novel condition creator based on our similarity-aware distribution sampling strategies to deliberately find novel and realistic inter action poses that are distinctive from the training set. Equipped with our metho d several baselines can be significantly improved beyond the SOTA on the HO3D an d DexYCB benchmarks. Our code will be released on https://github.com/hxwork/Hand Booster_Pytorch.
****
A-Teacher: Asymmetric Network for 3D Semi-Supervised Object Detection
Hanshi Wang, Zhipeng Zhang, Jin Gao, Weiming Hu; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14978-14987
This work proposes the first online asymmetric semi-supervised framework namely A-Teacher for LiDAR-based 3D object detection. Our motivation stems from the obs ervation that 1) existing symmetric teacher-student methods for semi-supervised 3D object detection have characterized simplicity but impede the distillation pe rformance between teacher and student because of the demand for an identical mod el structure and input data format. 2) The offline asymmetric methods with a com plex teacher model constructed differently can generate more precise pseudo-labe ls but is challenging to jointly optimize the teacher and student model. Consequ ently in this paper we devise a different path from the conventional paradigm wh ich can harness the capacity of a strong teacher while preserving the advantages of online teacher model updates. The essence is the proposed attention-based re finement model that can be seamlessly integrated into a vanilla teacher. The ref inement model works in the divide-and-conquer manner that respectively handles t hree challenging scenarios including 1) objects detected in the current timestam p but with suboptimal box quality 2) objects are missed in the current timestamp but are detected in past or future frames 3) objects are neglected in all frame s. It is worth noting that even while tackling these complex cases our model ret ains the efficiency of the online teacher-student semi-supervised framework. Exp erimental results on Waymo show that our method outperforms previous state-of-th e-art HSSDA for 4.7 on mAP (L1) while consuming fewer training resources.
****
GoMVS: Geometrically Consistent Cost Aggregation for Multi-View Stereo
Jiang Wu, Rui Li, Haofei Xu, Wenxun Zhao, Yu Zhu, Jinqiu Sun, Yanning Zhang; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20207-20216

Matching cost aggregation plays a fundamental role in learning-based multi-view stereo networks. However directly aggregating adjacent costs can lead to suboptimal results due to local geometric inconsistency. Related methods either seek selective aggregation or improve aggregated depth in the 2D space both are unable to handle geometric inconsistency in the cost volume effectively. In this paper we propose GoMVS to aggregate geometrically consistent costs yielding better utilization of adjacent geometries. More specifically we correspond and propagate adjacent costs to the reference pixel by leveraging the local geometric smoothness in conjunction with surface normals. We achieve this by the geometric consistent propagation (GCP) module. It computes the correspondence from the adjacent depth hypothesis space to the reference depth space using surface normals then uses the correspondence to propagate adjacent costs to the reference geometry followed by a convolution for aggregation. Our method achieves new state-of-the-art performance on DTU Tanks & Temple and ETH3D datasets. Notably our method ranks 1st on the Tanks & Temple Advanced benchmark. Code is available at https://github.com/Wuuu3511/GoMVS.
****
Evaluating Transferability in Retrieval Tasks: An Approach Using MMD and Kernel Methods
Mengyu Dai, Amir Hossein Raffiee, Aashish Jain, Joshua Correa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22390-22400
Retrieval tasks play central roles in real-world machine learning systems such as search engines recommender systems and retrieval-augmented generation (RAG). Achieving decent performance in these tasks often requires fine-tuning various pre-trained models on specific datasets and selecting the best candidate a process that can be both time and resource-consuming. To tackle the problem we introduce a novel and efficient method called RetMMD that leverages Maximum Mean Discrepancy (MMD) and kernel methods to assess the transferability of pretrained models in retrieval tasks. RetMMD is calculated on pretrained model and target dataset without any fine-tuning involved. Specifically given some query we quantify the distribution discrepancy between relevant and irrelevant document embeddings by estimating the similarities within their mappings in the fine-tuned embedding space through kernel method. This discrepancy is averaged over multiple queries taking into account the distribution characteristics of the target dataset. Experiments suggest that the proposed metric calculated on pre-trained models closely aligns with retrieval performance post-fine-tuning. The observation holds across a variety of datasets including image text and multi-modal domains indicating the potential of using MMD and kernel methods for transfer learning evaluation in retrieval scenarios. In addition we also design a way of evaluating dataset transferability for retrieval tasks with experimental results demonstrating the effectiveness of the proposed approach.
****
AnyScene: Customized Image Synthesis with Composited Foreground
Ruidong Chen, Lanjun Wang, Weizhi Nie, Yongdong Zhang, An-An Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8724-8733
Recent advancements in text-to-image technology have significantly advanced the field of image customization. Among various applications the task of customizing diverse scenes for user-specified composited elements holds great application value but has not been extensively explored. Addressing this gap we propose AnyScene a specialized framework designed to create varied scenes from composited foreground using textual prompts. AnyScene addresses the primary challenges inherent in existing methods particularly scene disharmony due to a lack of foreground semantic understanding and distortion of foreground elements. Specifically we develop a foreground injection module that guides a pre-trained diffusion model to generate cohesive scenes in visual harmony with the provided foreground. To enhance robust generation we implement a layout control strategy that prevents distortions of foreground elements. Furthermore an efficient image blending mechanism seamlessly reintegrates foreground details into the generated scenes producing

outputs with overall visual harmony and precise foreground details. In addition we propose a new benchmark and a series of quantitative metrics to evaluate this proposed image customization task. Extensive experimental results demonstrate the effectiveness of AnyScene which confirms its potential in various applications.
****
Training Generative Image Super-Resolution Models by Wavelet-Domain Losses Enables Better Control of Artifacts
Cansu Korkmaz, A. Murat Tekalp, Zafer Dogan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5926-5936
Super-resolution (SR) is an ill-posed inverse problem where the size of the set of feasible solutions that are consistent with a given low-resolution image is very large. Many algorithms have been proposed to find a "good" solution among the feasible solutions that strike a balance between fidelity and perceptual quality. Unfortunately all known methods generate artifacts and hallucinations while trying to reconstruct high-frequency (HF) image details. A fundamental question is: Can a model learn to distinguish genuine image details from artifacts? Although some recent works focused on the differentiation of details and artifacts this is a very challenging problem and a satisfactory solution is yet to be found. This paper shows that the characterization of genuine HF details versus artifacts can be better learned by training GAN-based SR models using wavelet-domain loss functions compared to RGB-domain or Fourier-space losses. Although wavelet-domain losses have been used in the literature before they have not been used in the context of the SR task. More specifically we train the discriminator only on the HF wavelet sub-bands instead of on RGB images and the generator is trained by a fidelity loss over wavelet subbands to make it sensitive to the scale and orientation of structures. Extensive experimental results demonstrate that our model achieves better perception-distortion trade-off according to multiple objective measures and visual evaluations.
****
Visual Objectification in Films: Towards a New AI Task for Video Interpretation
Julie Tores, Lucile Sassatelli, Hui-Yin Wu, Clement Bergman, Léa Andolfi, Victor Ecrement, Frédéric Precioso, Thierry Devars, Magali Guaresi, Virginie Julliard, Sarah Lecossais; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10864-10874
In film gender studies the concept of "male gaze" refers to the way the characters are portrayed on-screen as objects of desire rather than subjects. In this article we introduce a novel video-interpretation task to detect character objectification in films. The purpose is to reveal and quantify the usage of complex temporal patterns operated in cinema to produce the cognitive perception of objectification. We introduce the ObyGaze12 dataset made of 1914 movie clips densely annotated by experts for objectification concepts identified in film studies and psychology. We evaluate recent vision models show the feasibility of the task and where the challenges remain with concept bottleneck models. Our new dataset and code are made available to the community.
****
OMG-Seg: Is One Model Good Enough For All Segmentation?
Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27948-27959
In this work we address various segmentation tasks each traditionally tackled by distinct or partially unified models. We propose OMG-Seg One Model that is Good enough to efficiently and effectively handle all the segmentation tasks including image semantic instance and panoptic segmentation as well as their video counterparts open vocabulary settings prompt-driven interactive segmentation like SAM and video object segmentation. To our knowledge this is the first model to handle all these tasks in one model and achieve satisfactory performance. We show that OMG-Seg a transformer-based encoder-decoder architecture with task-specific queries and outputs can support over ten distinct segmentation tasks and yet significantly reduce computational and parameter overhead across various tasks and

datasets. We rigorously evaluate the inter-task influences and correlations duri
ng co-training. Code and models are available at https://github.com/lxtGH/OMG-Se
g.
****
BiTT: Bi-directional Texture Reconstruction of Interacting Two Hands from a Sing
le Image
Minje Kim, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 10726-10735
Creating personalized hand avatars is important to offer a realistic experience
to users on AR / VR platforms. While most prior studies focused on reconstructin
g 3D hand shapes some recent work has tackled the reconstruction of hand texture
s on top of shapes. However these methods are often limited to capturing pixels
on the visible side of a hand requiring diverse views of the hand in a video or
multiple images as input. In this paper we propose a novel method BiTT(Bi-direct
ional Texture reconstruction of Two hands) which is the first end-to-end train-
able method for relightable pose-free texture reconstruction of two interacting
hands taking only a single RGB image by three novel components: 1) bi-directiona
l (left ? right) texture reconstruction using the texture symmetry of left / rig
ht hands 2) utilizing a texture parametric model for hand texture recovery and 3
) the overall coarse-to-fine stage pipeline for reconstructing personalized text
ure of two interacting hands. BiTT first estimates the scene light condition and
 albedo image from an input image then reconstructs the texture of both hands th
rough the texture parametric model and bi-directional texture reconstructor. In
experiments using InterHand2.6M and RGB2Hands datasets our method significantly
outperforms state-of-the-art hand texture reconstruction methods quantitatively
and qualitatively. The code is available at https://github.com/ yunminjin2/BiTT.
****
DetCLIPv3: Towards Versatile Generative Open-vocabulary Object Detection
Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li
, Dan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2024, pp. 27391-27401
Existing open-vocabulary object detectors typically require a predefined set of
categories from users significantly confining their application scenarios. In th
is paper we introduce DetCLIPv3 a high-performing detector that excels not only
at both open-vocabulary object detection but also generating hierarchical labels
 for detected objects. DetCLIPv3 is characterized by three core designs: 1. Vers
atile model architecture: we derive a robust open-set detection framework which
is further empowered with generation ability via the integration of a caption he
ad. 2. High information density data: we develop an auto-annotation pipeline lev
eraging visual large language model to refine captions for large-scale image-tex
t pairs providing rich multi-granular object labels to enhance the training. 3.
Efficient training strategy: we employ a pre-training stage with low-resolution
inputs that enables the object captioner to efficiently learn a broad spectrum o
f visual concepts from extensive image-text paired data. This is followed by a f
ine-tuning stage that leverages a small number of high-resolution samples to fur
ther enhance detection performance. With these effective designs DetCLIPv3 demon
strates superior open-vocabulary detection performance e.g. our Swin-T backbone
model achieves a notable 47.0 zero-shot fixed AP on the LVIS minival benchmark o
utperforming GLIPv2 GroundingDINO and DetCLIPv2 by 18.0/19.6/6.6 AP respectively
. DetCLIPv3 also achieves a state-of-the-art 19.7 AP in dense captioning task on
 VG dataset showcasing its strong generative capability.
****
UVEB: A Large-scale Benchmark and Baseline Towards Real-World Underwater Video E
nhancement
Yaofeng Xie, Lingwei Kong, Kai Chen, Ziqiang Zheng, Xiao Yu, Zhibin Yu, Bing Zhe
ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2024, pp. 22358-22367
Learning-based underwater image enhancement (UIE) methods have made great progre
ss. However the lack of large-scale and high-quality paired training samples has
 become the main bottleneck hindering the development of UIE. The inter-frame in

formation in underwater videos can accelerate or optimize the UIE process. Thus we constructed the first large-scale high-resolution underwater video enhancement benchmark (UVEB) to promote the development of underwater vision.It contains 1308 pairs of video sequences and more than 453000 high-resolution with 38% Ultra-High-Definition (UHD) 4K frame pairs. UVEB comes from multiple countries containing various scenes and video degradation types to adapt to diverse and complex underwater environments. We also propose the first supervised underwater video enhancement method UVE-Net. UVE-Net converts the current frame information into convolutional kernels and passes them to adjacent frames for efficient inter-frame information exchange. By fully utilizing the redundant degraded information of underwater videos UVE-Net completes video enhancement better. Experiments show the effective network design and good performance of UVE-Net.
****
Learning to Localize Objects Improves Spatial Reasoning in Visual-LLMs
Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, Tsung-Yu Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12977-12987
Integration of Large Language Models (LLMs) into visual domain tasks resulting in visual-LLMs (V-LLMs) has enabled exceptional performance in vision-language tasks particularly for visual question answering (VQA). However existing V-LLMs (e.g. BLIP-2 LLaVA) demonstrate weak spatial reasoning and localization awareness. Despite generating highly descriptive and elaborate textual answers these models fail at simple tasks like distinguishing a left vs right location. In this work we explore how image-space coordinate based instruction fine-tuning objectives could inject spatial awareness into V-LLMs. We discover optimal coordinate representations data-efficient instruction fine-tuning objectives and pseudo-data generation strategies that lead to improved spatial awareness in V-LLMs. Additionally our resulting model improves VQA across image and video domains reduces undesired hallucination and generates better contextual object descriptions. Experiments across 5 vision-language tasks involving 14 different datasets establish the clear performance improvements achieved by our proposed framework.
****
Monocular Identity-Conditioned Facial Reflectance Reconstruction
Xingyu Ren, Jiankang Deng, Yuhao Cheng, Jia Guo, Chao Ma, Yichao Yan, Wenhan Zhu, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 885-895
Recent 3D face reconstruction methods have made remarkable advancements yet there remain huge challenges in monocular high-quality facial reflectance reconstruction. Existing methods rely on a large amount of light-stage captured data to learn facial reflectance models. However the lack of subject diversity poses challenges in achieving good generalization and widespread applicability. In this paper we learn the reflectance prior in image space rather than UV space and present a framework named ID2Reflectance. Our framework can directly estimate the reflectance maps of a single image while using limited reflectance data for training. Our key insight is that reflectance data shares facial structures with RGB faces which enables obtaining expressive facial prior from inexpensive RGB data thus reducing the dependency on reflectance data. We first learn a high-quality prior for facial reflectance. Specifically we pretrain multi-domain facial feature codebooks and design a codebook fusion method to align the reflectance and RGB domains. Then we propose an identity-conditioned swapping module that injects facial identity from the target image into the pre-trained auto-encoder to modify the identity of the source reflectance image. Finally we stitch multi-view swapped reflectance images to obtain renderable assets. Extensive experiments demonstrate that our method exhibits excellent generalization capability and achieves state-of-the-art facial reflectance reconstruction results for in-the-wild faces.
****
C3: High-Performance and Low-Complexity Neural Compression from a Single Image or Video
Hyunjik Kim, Matthias Bauer, Lucas Theis, Jonathan Richard Schwarz, Emilien Dupont; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn

ition (CVPR), 2024, pp. 9347-9358
Most neural compression models are trained on large datasets of images or videos in order to generalize to unseen data. Such generalization typically requires large and expressive architectures with a high decoding complexity. Here we introduce C3 a neural compression method with strong rate-distortion (RD) performance that instead overfits a small model to each image or video separately. The resulting decoding complexity of C3 can be an order of magnitude lower than neural baselines with similar RD performance. C3 builds on COOL-CHIC [Ladune et al 2023] and makes several simple and effective improvements for images. We further develop new methodology to apply C3 to videos. On the CLIC2020 image benchmark we match the RD performance of VTM the reference implementation of the H.266 codec with less than3k MACs/pixel for decoding. On the UVG video benchmark we match the RD performance of the Video Compression Transformer [Mentzer er al 2022] a well-established neural video codec with less than 5k MACs/pixel for decoding.
****
Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors
Nicolae-C?t?lin Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15984-15995
We propose an efficient abnormal event detection model based on a lightweight masked auto-encoder (AE) applied at the video frame level. The novelty of the proposed model is threefold. First we introduce an approach to weight tokens based on motion gradients thus shifting the focus from the static background scene to the foreground objects. Second we integrate a teacher decoder and a student decoder into our architecture leveraging the discrepancy between the outputs given by the two decoders to improve anomaly detection. Third we generate synthetic abnormal events to augment the training videos and task the masked AE model to jointly reconstruct the original frames (without anomalies) and the corresponding pixel-level anomaly maps. Our design leads to an efficient and effective model as demonstrated by the extensive experiments carried out on four benchmarks: Avenue ShanghaiTech UBnormal and UCSD Ped2. The empirical results show that our model achieves an excellent trade-off between speed and accuracy obtaining competitive AUC scores while processing 1655 FPS. Hence our model is between 8 and 70 times faster than competing methods. We also conduct an ablation study to justify our design. Our code is freely available at: https://github.com/ristea/aed-mae.
****
Revisiting Non-Autoregressive Transformers for Efficient Image Synthesis
Zanlin Ni, Yulin Wang, Renping Zhou, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Shiji Song, Yuan Yao, Gao Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7007-7016
The field of image synthesis is currently flourishing due to the advancements in diffusion models. While diffusion models have been successful their computational intensity has prompted the pursuit of more efficient alternatives. As a representative work non-autoregressive Transformers (NATs) have been recognized for their rapid generation. However a major drawback of these models is their inferior performance compared to diffusion models. In this paper we aim to re-evaluate the full potential of NATs by revisiting the design of their training and inference strategies. Specifically we identify the complexities in properly configuring these strategies and indicate the possible sub-optimality in existing heuristic-driven designs. Recognizing this we propose to go beyond existing methods by directly solving the optimal strategies in an automatic framework. The resulting method named AutoNAT advances the performance boundaries of NATs notably and is able to perform comparably with the latest diffusion models with a significantly reduced inference cost. The effectiveness of AutoNAT is comprehensively validated on four benchmark datasets i.e. ImageNet-256 & 512 MS-COCO and CC3M. Code and pre-trained models will be available at https://github.com/LeapLabTHU/ImprovedNAT.
****
Distilling Vision-Language Models on Millions of Videos
Yue Zhao, Long Zhao, Xingyi Zhou, Jialin Wu, Chun-Te Chu, Hui Miao, Florian Schr

off, Hartwig Adam, Ting Liu, Boqing Gong, Philipp Krahenbuhl, Liangzhe Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13106-13116

The recent advance in vision-language models is largely attributed to the abundance of image-text data. We aim to replicate this success for video-language models but there simply is not enough human-curated video-text data available. We thus resort to fine-tuning a video-language model from a strong image-language baseline with synthesized instructional data. The resulting video model by video-instruction-tuning (VIIT) is then used to auto-label millions of videos to generate high-quality captions. We show the adapted video-language model performs well on a wide range of video-language benchmarks. For instance it surpasses the best prior result on open-ended NExT-QA by2.8%. Besides our model generates detailed descriptions for previously unseen videos which provide better textual supervision than existing methods. Experiments show that a video-language dual-encoder model contrastively trained on these auto-generated captions is 3.8% better than the strongest baseline that also leverages vision-language models. Our best model outperforms state-of-the-art methods on MSR-VTT zero-shot text-to-video retrieval by 6%. As a side product we generate the largest video caption dataset to date.

****

ANIM: Accurate Neural Implicit Model for Human Reconstruction from a single RGB-D Image

Marco Pesavento, Yuanlu Xu, Nikolaos Sarafianos, Robert Maier, Ziyan Wang, Chun-Han Yao, Marco Volino, Edmond Boyer, Adrian Hilton, Tony Tung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5448-5458

Recent progress in human shape learning shows that neural implicit models are effective in generating 3D human surfaces from limited number of views and even from a single RGB image. However existing monocular approaches still struggle to recover fine geometric details such as face hands or cloth wrinkles. They are also easily prone to depth ambiguities that result in distorted geometries along the camera optical axis. In this paper we explore the benefits of incorporating depth observations in the reconstruction process by introducing ANIM a novel method that reconstructs arbitrary 3D human shapes from single-view RGB-D images with an unprecedented level of accuracy. Our model learns geometric details from both multi-resolution pixel-aligned and voxel-aligned features to leverage depth information and enable spatial relationships mitigating depth ambiguities. We further enhance the quality of the reconstructed shape by introducing a depth-supervision strategy which improves the accuracy of the signed distance field estimation of points that lie on the reconstructed surface. Experiments demonstrate that ANIM outperforms state-of-the-art works that use RGB surface normals point cloud or RGB-D data as input. In addition we introduce ANIM-Real a new multi-modal dataset comprising high-quality scans paired with consumer-grade RGB-D camera and our protocol to fine-tune ANIM enabling high-quality reconstruction from real-world human capture.

****

Real-Time Simulated Avatar from Head-Mounted Sensors

Zhengyi Luo, Jinkun Cao, Rawal Khirodkar, Alexander Winkler, Kris Kitani, Weipeng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 571-581

We present SimXR a method for controlling a simulated avatar from information (headset pose and cameras) obtained from AR / VR headsets. Due to the challenging viewpoint of head-mounted cameras the human body is often clipped out of view making traditional image-based egocentric pose estimation challenging. On the other hand headset poses provide valuable information about overall body motion but lack fine-grained details about the hands and feet. To synergize headset poses with cameras we control a humanoid to track headset movement while analyzing input images to decide body movement. When body parts are seen the movements of hands and feet will be guided by the images; when unseen the laws of physics guide the controller to generate plausible motion. We design an end-to-end method that

does not rely on any intermediate representations and learns to directly map from images and headset poses to humanoid control signals. To train our method we also propose a large-scale synthetic dataset created using camera configurations compatible with a commercially available VR headset (Quest 2) and show promising results on real-world captures. To demonstrate the applicability of our framework we also test it on an AR headset with a forward-facing camera.
****
Discovering Syntactic Interaction Clues for Human-Object Interaction Detection
Jinguo Luo, Weihong Ren, Weibo Jiang, Xi'ai Chen, Qiang Wang, Zhi Han, Honghai Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28212-28222
Recently Vision-Language Model (VLM) has greatly advanced the Human-Object Interaction (HOI) detection. The existing VLM-based HOI detectors typically adopt a hand-crafted template (e.g. a photo of a person [action] a/an [object]) to acquire text knowledge through the VLM text encoder. However such approaches only encoding the action-specific text prompts in vocabulary level may suffer from learning ambiguity without exploring the fine-grained clues from the perspective of interaction context. In this paper we propose a novel method to discover Syntactic Interaction Clues for HOI detection (SICHOI) by using VLM. Specifically we first investigate what are the essential elements for an interaction context and then establish a syntactic interaction bank from three levels: spatial relationship action-oriented posture and situational condition. Further to align visual features with the syntactic interaction bank we adopt a multi-view extractor to jointly aggregate visual features from instance interaction and image levels accordingly. In addition we also introduce a dual cross-attention decoder to perform context propagation between text knowledge and visual features thereby enhancing the HOI detection. Experimental results demonstrate that our proposed method achieves state-of-the-art performance on HICO-DET and V-COCO.
****
Inter-X: Towards Versatile Human-Human Interaction Analysis
Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22260-22271
The analysis of the ubiquitous human-human interactions is pivotal for understanding humans as social beings. Existing human-human interaction datasets typically suffer from inaccurate body motions lack of hand gestures and fine-grained textual descriptions. To better perceive and generate human-human interactions we propose Inter-X a currently largest human-human interaction dataset with accurate body movements and diverse interaction patterns together with detailed hand gestures. The dataset includes 11K interaction sequences and more than 8.1M frames. We also equip Inter-X with versatile annotations of more than 34K fine-grained human part-level textual descriptions semantic interaction categories interaction order and the relationship and personality of the subjects. Based on the elaborate annotations we propose a unified benchmark composed of 4 categories of downstream tasks from both the perceptual and generative directions. Extensive experiments and comprehensive analysis show that Inter-X serves as a testbed for promoting the development of versatile human-human interaction analysis. Our dataset and benchmark will be publicly available for research purposes.
****
Generalized Predictive Model for Autonomous Driving
Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, Hongyang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14662-14672
In this paper we introduce the first large-scale video prediction model in the autonomous driving discipline. To eliminate the restriction of high-cost data collection and empower the generalization ability of our model we acquire massive data from the web and pair it with diverse and high-quality text descriptions. The resultant dataset accumulates over 2000 hours of driving videos spanning areas

all over the world with diverse weather conditions and traffic scenarios. Inheriting the merits from recent latent diffusion models our model dubbed GenAD handles the challenging dynamics in driving scenes with novel temporal reasoning blocks. We showcase that it can generalize to various unseen driving datasets in a zero-shot manner surpassing general or driving-specific video prediction counterparts. Furthermore GenAD can be adapted into an action-conditioned prediction model or a motion planner holding great potential for real-world driving applications.
****
FACT: Frame-Action Cross-Attention Temporal Modeling for Efficient Action Segmentation
Zijia Lu, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18175-18185
We study supervised action segmentation whose goal is to predict framewise action labels of a video. To capture temporal dependencies over long horizons prior works either improve framewise features with transformer or refine framewise predictions with learned action features. However they are computationally costly and ignore that frame and action features contain complimentary information which can be leveraged to enhance both features and improve temporal modeling. Therefore we propose an efficient Frame-Action Cross-attention Temporal modeling (FACT) framework that performs temporal modeling with frame and action features in parallel and leverage this parallelism to achieve iterative bidirectional information transfer between the features and refine them. FACT network contains (i) a frame branch to learn frame-level information with convolutions and frame features (ii) an action branch to learn action-level dependencies with transformers and action tokens and (iii) cross-attentions to allow communication between the two branches. We also propose a new matching loss to ensure each action token uniquely encodes an action segment thus better captures its semantics. Thanks to our architecture we can also leverage textual transcripts of videos to help action segmentation. We evaluate FACT on four video datasets (two egocentric and two third-person) for action segmentation with and without transcripts showing that it significantly improves the state-of-the-art accuracy while enjoys lower computational cost (3 times faster) than existing transformer-based methods
****
Test-Time Zero-Shot Temporal Action Localization
Benedetta Liberatori, Alessandro Conti, Paolo Rota, Yiming Wang, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18720-18729
Zero-Shot Temporal Action Localization (ZS-TAL) seeks to identify and locate actions in untrimmed videos unseen during training. Existing ZS-TAL methods involve fine-tuning a model on a large amount of annotated training data. While effective training-based ZS-TAL approaches assume the availability of labeled data for supervised learning which can be impractical in some applications. Furthermore the training process naturally induces a domain bias into the learned model which may adversely affect the model's generalization ability to arbitrary videos. These considerations prompt us to approach the ZS-TAL problem from a radically novel perspective relaxing the requirement for training data. To this aim we introduce a novel method that performs Test-Time adaptation for Temporal Action Localization (T3AL). In a nutshell T3AL adapts a pre-trained Vision and Language Model (VLM). T3AL operates in three steps. First a video-level pseudo-label of the action category is computed by aggregating information from the entire video. Then action localization is performed adopting a novel procedure inspired by self-supervised learning. Finally frame-level textual descriptions extracted with a state-of-the-art captioning model are employed for refining the action region proposals. We validate the effectiveness of T 3AL by conducting experiments on the THUMOS14 and the ActivityNet-v1.3 datasets. Our results demonstrate that T3AL significantly outperforms zero-shot baselines based on state-of-the-art VLMs confirming the benefit of a test-time adaptation approach.
****
AM-RADIO: Agglomerative Vision Foundation Model Reduce All Domains Into One

Mike Ranzinger, Greg Heinrich, Jan Kautz, Pavlo Molchanov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12490-12500
A handful of visual foundation models (VFMs) have recently emerged as the backbones for numerous downstream tasks. VFMs like CLIP DINOv2 SAM are trained with distinct objectives exhibiting unique characteristics for various downstream tasks. We find that despite their conceptual differences these models can be effectively merged into a unified model through multi-teacher distillation. We name this approach AM-RADIO (Agglomerative Model -- Reduce All Domains Into One). This integrative approach not only surpasses the performance of individual teacher models but also amalgamates their distinctive features such as zero-shot vision-language comprehension detailed pixel-level understanding and open vocabulary segmentation capabilities. Additionally in pursuit of the most hardware-efficient backbone we evaluated numerous architectures in our multi-teacher distillation pipeline using the same training recipe. This led to the development of a novel architecture (E-RADIO) that exceeds the performance of its predecessors and is at least 6x faster than the teacher models at matched resolution. Our comprehensive benchmarking process covers downstream tasks including ImageNet classification semantic segmentation linear probing COCO object detection and integration into LLaVa-1.5.
****
MaskClustering: View Consensus based Mask Graph Clustering for Open-Vocabulary 3D Instance Segmentation
Mi Yan, Jiazhao Zhang, Yan Zhu, He Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28274-28284
Open-vocabulary 3D instance segmentation is cutting-edge for its ability to segment 3D instances without predefined categories. However progress in 3D lags behind its 2D counterpart due to limited annotated 3D data. To address this recent works first generate 2D open-vocabulary masks through 2D models and then merge them into 3D instances based on metrics calculated between two neighboring frames. In contrast to these local metrics we propose a novel metric view consensus rate to enhance the utilization of multi-view observations. The key insight is that two 2D masks should be deemed part of the same 3D instance if a significant number of other 2D masks from different views contain both these two masks. Using this metric as edge weight we construct a global mask graph where each mask is a node. Through iterative clustering of masks showing high view consensus we generate a series of clusters each representing a distinct 3D instance. Notably our model is training-free. Through extensive experiments on publicly available datasets including ScanNet++ ScanNet200 and MatterPort3D we demonstrate that our method achieves state-of-the-art performance in open-vocabulary 3D instance segmentation. Our project page is at \href https://pku-epic.github.io/MaskClustering/ https://pku-epic.github.io/MaskClustering .
****
Seamless Human Motion Composition with Blended Positional Encodings
German Barquero, Sergio Escalera, Cristina Palmero; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 457-469
Conditional human motion generation is an important topic with many applications in virtual reality gaming and robotics. While prior works have focused on generating motion guided by text music or scenes these typically result in isolated motions confined to short durations. Instead we address the generation of long continuous sequences guided by a series of varying textual descriptions. In this context we introduce FlowMDM the first diffusion-based model that generates seamless Human Motion Compositions (HMC) without any postprocessing or redundant denoising steps. For this we introduce the Blended Positional Encodings a technique that leverages both absolute and relative positional encodings in the denoising chain. More specifically global motion coherence is recovered at the absolute stage whereas smooth and realistic transitions are built at the relative stage. As a result we achieve state-of-the-art results in terms of accuracy realism and smoothness on the Babel and HumanML3D datasets. FlowMDM excels when trained with only a single description per motion sequence thanks to its Pose-Centric Cross-A

Ttention which makes it robust against varying text descriptions at inference time. Finally to address the limitations of existing HMC metrics we propose two new metrics: the Peak Jerk and the Area Under the Jerk to detect abrupt transitions.
****
PeerAiD: Improving Adversarial Distillation from a Specialized Peer Tutor
Jaewon Jung, Hongsun Jang, Jaeyong Song, Jinho Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24482-24491
Adversarial robustness of the neural network is a significant concern when it is applied to security-critical domains. In this situation adversarial distillation is a promising option which aims to distill the robustness of the teacher network to improve the robustness of a small student network. Previous works pretrain the teacher network to make it robust against the adversarial examples aimed at itself. However the adversarial examples are dependent on the parameters of the target network. The fixed teacher network inevitably degrades its robustness against the unseen transferred adversarial examples which target the parameters of the student network in the adversarial distillation process. We propose PeerAiD to make a peer network learn the adversarial examples of the student network instead of adversarial examples aimed at itself. PeerAiD is an adversarial distillation that trains the peer network and the student network simultaneously in order to specialize the peer network for defending the student network. We observe that such peer networks surpass the robustness of the pretrained robust teacher model against adversarial examples aimed at the student network. With this peer network and adversarial distillation PeerAiD achieves significantly higher robustness of the student network with AutoAttack (AA) accuracy by up to 1.66%p and improves the natural accuracy of the student network by up to 4.72%p with ResNet-18 on TinyImageNet dataset. Code is available at https://github.com/jaewonalive/PeerAiD.
****
Scaling Laws for Data Filtering-- Data Curation cannot be Compute Agnostic
Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, J. Zico Kolter; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22702-22711
Vision-language models (VLMs) are trained for thousands of GPU hours on carefully selected subsets of massive web scrapes. For instance the LAION public dataset retained only about 10 percent of the total crawled data. In recent times data curation has gained prominence with several works developing strategies to retain high-quality subsets of raw scraped data. However these strategies are typically developed agnostic to the available compute for training. In this paper we demonstrate that making filtering decisions independent of training compute is often suboptimal: well-curated data rapidly loses its utility when repeated eventually decreasing below the utility of unseen but lower-quality data. While past research in neural scaling laws has considered web data to be homogenous real data is not. Our work bridges this important gap in the literature by developing scaling laws that characterize the differing utility of various data subsets and accounting for how this diminishes for a data point at its nth repetition. Our key message is that data curation can not be agnostic of the total compute a model will be trained for. Even without ever jointly training on multiple data buckets our scaling laws enable us to estimate model performance under this dynamic trade-off between quality and repetition. This allows us to curate the best possible pool for achieving top performance on Datacomp at various compute budgets carving out a pareto-frontier for data curation.
****
FastMAC: Stochastic Spectral Sampling of Correspondence Graph
Yifei Zhang, Hao Zhao, Hongyang Li, Siheng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17857-17867
3D correspondence i.e. a pair of 3D points is a fundamental concept in computer vision. A set of 3D correspondences when equipped with compatibility edges forms a correspondence graph. This graph is a critical component in several state-of-

the-art 3D point cloud registration approaches e.g. the one based on maximal cliques (MAC). However its properties have not been well understood. So we present the first study that introduces graph signal processing into the domain of correspondence graph. We exploit the generalized degree signal on correspondence graph and pursue sampling strategies that preserve high-frequency components of this signal. To address time-consuming singular value decomposition in deterministic sampling we resort to a stochastic approximate sampling strategy. As such the core of our method is the stochastic spectral sampling of correspondence graph. As an application we build a complete 3D registration algorithm termed as FastMAC that reaches real-time speed while leading to little to none performance drop. Through extensive experiments we validate that FastMAC works for both indoor and outdoor benchmarks. For example FastMAC can accelerate MAC by 80 times while maintaining high registration success rate on KITTI. Codes are publicly available at https://github.com/Forrest-110/FastMAC.
****
FedUV: Uniformity and Variance for Heterogeneous Federated Learning
Ha Min Son, Moon-Hyun Kim, Tai-Myoung Chung, Chao Huang, Xin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5863-5872
Federated learning is a promising framework to train neural networks with widely distributed data. However performance degrades heavily with heterogeneously distributed data. Recent work has shown this is due to the final layer of the network being most prone to local bias some finding success freezing the final layer as an orthogonal classifier. We investigate the training dynamics of the classifier by applying SVD to the weights motivated by the observation that freezing weights results in constant singular values. We find that there are differences when training in IID and non-IID settings. Based on this finding we introduce two regularization terms for local training to continuously emulate IID settings: (1) variance in the dimension-wise probability distribution of the classifier and (2) hyperspherical uniformity of representations of the encoder. These regularizations promote local models to act as if it were in an IID setting regardless of the local data distribution thus offsetting proneness to bias while being flexible to the data. On extensive experiments in both label-shift and feature-shift settings we verify that our method achieves highest performance by a large margin especially in highly non-IID cases in addition to being scalable to larger models and datasets.
****
FedSOL: Stabilized Orthogonal Learning with Proximal Restrictions in Federated Learning
Gihun Lee, Minchan Jeong, Sangmook Kim, Jaehoon Oh, Se-Young Yun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12512-12522
Federated Learning (FL) aggregates locally trained models from individual clients to construct a global model. While FL enables learning a model with data privacy it often suffers from significant performance degradation when clients have heterogeneous data distributions. This data heterogeneity causes the model to forget the global knowledge acquired from previously sampled clients after being trained on local datasets. Although the introduction of proximal objectives in local updates helps to preserve global knowledge it can also hinder local learning by interfering with local objectives. Inspired by Continual Learning (CL) we adopt an orthogonal learning strategy to balance these two conflicting objectives. However we observe that directly negating the proximal gradient in the local gradient significantly undermines local learning. To address the problem we propose a novel method Federated Stabilized Orthogonal Learning (FedSOL). FedSOL is designed to identify gradients of local objectives that are inherently orthogonal to directions affecting the proximal objective. Specifically FedSOL targets parameter regions where learning on the local objective is minimally influenced by proximal weight perturbations. Our experiments demonstrate that FedSOL consistently achieves state-of-the-art performance across various scenarios.
****

GAvatar: Animatable 3D Gaussian Avatars with Implicit Mesh Learning
Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, Umar Iqbal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 896-905
Gaussian splatting has emerged as a powerful 3D representation that harnesses the advantages of both explicit (mesh) and implicit (NeRF) 3D representations. In this paper we seek to leverage Gaussian splatting to generate realistic animatable avatars from textual descriptions addressing the limitations (e.g. efficiency and flexibility) imposed by mesh or NeRF-based representations. However a naive application of Gaussian splatting cannot generate high-quality animatable avatars and suffers from learning instability; it also cannot capture fine avatar geometries and often leads to degenerate body parts. To tackle these problems we first propose a primitive-based 3D Gaussian representation where Gaussians are defined inside pose-driven primitives to facilitate animations. Second to stabilize and amortize the learning of millions of Gaussians we propose to use implicit neural fields to predict the Gaussian attributes (e.g. colors). Finally to capture fine avatar geometries and extract detailed meshes we propose a novel SDF-based implicit mesh learning approach for 3D Gaussians that regularizes the underlying geometries and extracts highly detailed textured meshes. Our proposed method GAvatar enables the large-scale generation of diverse animatable avatars using only text prompts. GAvatar significantly surpasses existing methods in terms of both appearance and geometry quality and achieves extremely fast rendering (100 fps) at 1K resolution.
****
Beyond Average: Individualized Visual Scanpath Prediction
Xianyu Chen, Ming Jiang, Qi Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25420-25431
Understanding how attention varies across individuals has significant scientific and societal impacts. However existing visual scanpath models treat attention uniformly neglecting individual differences. To bridge this gap this paper focuses on individualized scanpath prediction (ISP) a new attention modeling task that aims to accurately predict how different individuals shift their attention in diverse visual tasks. It proposes an ISP method featuring three novel technical components: (1) an observer encoder to characterize and integrate an observer's unique attention traits (2) an observer-centric feature integration approach that holistically combines visual features task guidance and observer-specific characteristics and (3) an adaptive fixation prioritization mechanism that refines scanpath predictions by dynamically prioritizing semantic feature maps based on individual observers' attention traits. These novel components allow scanpath models to effectively address the attention variations across different observers. Our method is generally applicable to different datasets model architectures and visual tasks offering a comprehensive tool for transforming general scanpath models into individualized ones. Comprehensive evaluations using value-based and ranking-based metrics verify the method's effectiveness and generalizability.
****
A Category Agnostic Model for Visual Rearrangment
Yuyi Liu, Xinhang Song, Weijie Li, Xiaohan Wang, Shuqiang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16457-16466
This paper presents a novel category agnostic model for visual rearrangement task which can help an embodied agent to physically recover the shuffled scene configuration without any category concepts to the goal configuration. Previous methods usually follow a similar architecture completing the rearrangement task by aligning the scene changes of the goal and shuffled configuration according to the semantic scene graphs. However constructing scene graphs requires the inference of category labels which not only causes the accuracy drop of the entire task but also limits the application in real world scenario. In this paper we delve deep into the essence of visual rearrangement task and focus on the two most essential issues scene change detection and scene change matching. We utilize the movement and the protrusion of point cloud to accurately identify the scene change

s and match these changes depending on the similarity of category agnostic appea rance feature. Moreover to assist the agent to explore the environment more effi ciently and comprehensively we propose a closer-aligned-retrace exploration poli cy aiming to observe more details of the scene at a closer distance. We conduct extensive experiments on AI2THOR Rearrangement Challenge based on RoomR dataset and a new multi-room multi-instance dataset MrMiR collected by us. The experimen tal results demonstrate the effectiveness of our proposed method.
****
Grounding Everything: Emerging Localization Properties in Vision-Language Transf ormers
Walid Bousselham, Felix Petersen, Vittorio Ferrari, Hilde Kuehne; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024 , pp. 3828-3837
Vision-language foundation models have shown remarkable performance in various z ero-shot settings such as image retrieval classification or captioning. But so f ar those models seem to fall behind when it comes to zero-shot localization of r eferential expressions and objects in images. As a result they need to be fine-t uned for this task. In this paper we show that pretrained vision-language (VL) m odels allow for zero-shot open-vocabulary object localization without any fine-t uning. To leverage those capabilities we propose a Grounding Everything Module (GEM) that generalizes the idea of value-value attention introduced by CLIPSurger y to a self-self attention path. We show that the concept of self-self attention corresponds to clustering thus enforcing groups of tokens arising from the same object to be similar while preserving the alignment with the language space. To further guide the group formation we propose a set of regularizations that allo ws the model to finally generalize across datasets and backbones. We evaluate th e proposed GEM framework on various benchmark tasks and datasets for semantic se gmentation. GEM not only outperforms other training-free open-vocabulary localiz ation methods but also achieves state-of-the-art results on the recently propose d OpenImagesV7 large-scale segmentation benchmark. Code is available at https:// github.com/WalBouss/GEM
****
Seeing Motion at Nighttime with an Event Camera
Haoyue Liu, Shihan Peng, Lin Zhu, Yi Chang, Hanyu Zhou, Luxin Yan; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 25648-25658
We focus on a very challenging task: imaging at nighttime dynamic scenes. Most p revious methods rely on the low-light enhancement of a conventional RGB camera. However they would inevitably face a dilemma between the long exposure time of n ighttime and the motion blur of dynamic scenes. Event cameras react to dynamic c hanges with higher temporal resolution (microsecond) and higher dynamic range (1 20dB) offering an alternative solution. In this work we present a novel nighttim e dynamic imaging method with an event camera. Specifically we discover that the event at nighttime exhibits temporal trailing characteristics and spatial non-s tationary distribution. Consequently we propose a nighttime event reconstruction network (NER-Net) which mainly includes a learnable event timestamps calibratio n module (LETC) to align the temporal trailing events and a non-uniform illumina tion aware module (NIAM) to stabilize the spatiotemporal distribution of events. Moreover we construct a paired real low-light event dataset (RLED) through a co -axial imaging system including 64200 spatially and temporally aligned image GTs and low-light events. Extensive experiments demonstrate that the proposed metho d outperforms state-of-the-art methods in terms of visual quality and generaliza tion ability on real-world nighttime datasets. The project are available at: htt ps://github.com/Liu-haoyue/NER-Net.
****
Representing Part-Whole Hierarchies in Foundation Models by Learning Localizabil ity Composability and Decomposability from Anatomy via Self Supervision
Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, Jianming Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 024, pp. 11269-11281

Humans effortlessly interpret images by parsing them into part-whole hierarchies
; deep learning excels in learning multi-level feature spaces but they often lac
k explicit coding of part-whole relations a prominent property of medical imagin
g. To overcome this limitation we introduce Adam-v2 a new self-supervised learni
ng framework extending Adam [68] by explicitly incorporating part-whole hierarch
ies into its learning objectives through three key branches: (1) Localizability
acquiring discriminative representations to distinguish different anatomical pat
terns; (2) Composability learning each anatomical structure in a parts-to-whole
manner; and (3) Decomposability comprehending each anatomical structure in a who
le-to-parts manner. Experimental results across 10 tasks compared to 11 baseline
s in zero-shot few-shot transfer and full fine-tuning settings showcase Adam-v2'
s superior performance over large-scale medical models and existing SSL methods
across diverse downstream tasks. The higher generality and robustness of Adam-v2
's representations originate from its explicit construction of hierarchies for d
istinct anatomical structures from unlabeled medical images. Adam-v2 preserves a
 semantic balance of anatomical diversity and harmony in its embedding yielding
representations that are both generic and semantically meaningful yet overlooked
 in existing SSL methods. All code and pretrained models are available at GitHub
.com/JLiangLab/Eden.
****
Efficient Test-Time Adaptation of Vision-Language Models
Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, Eric Xing; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2024, pp. 14162-14171
Test-time adaptation with pre-trained vision-language models has attracted incre
asing attention for tackling distribution shifts during the test time. Though pr
ior studies have achieved very promising performance they involve intensive comp
utation which is severely unaligned with test-time adaptation. We design TDA a t
raining-free dynamic adapter that enables effective and efficient test-time adap
tation with vision-language models. TDA works with a lightweight key-value cache
 that maintains a dynamic queue with few-shot pseudo labels as values and the co
rresponding test-sample features as keys. Leveraging the key-value cache TDA all
ows adapting to test data gradually via progressive pseudo label refinement whic
h is super-efficient without incurring any backpropagation. In addition we intro
duce negative pseudo labeling that alleviates the adverse impact of pseudo label
 noises by assigning pseudo labels to certain negative classes when the model is
 uncertain about its pseudo label predictions. Extensive experiments over two be
nchmarks demonstrate TDA's superior effectiveness and efficiency as compared wit
h the state-of-the-art. The code has been released in https://kdiaaa.github.io/t
da/.
****
Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs
Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, Saining Xie; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 9568-9578
Is vision good enough for language? Recent advancements in multimodal models pri
marily stem from the powerful reasoning abilities of large language models (LLMs
). However the visual component typically depends only on the instance-level con
trastive language-image pre-training (CLIP). Our research reveals that the visua
l capabilities in recent MultiModal LLMs (MLLMs) still exhibit systematic shortc
omings. To understand the roots of these errors we explore the gap between the v
isual embedding space of CLIP and vision-only self-supervised learning. We ident
ify "CLIP-blind pairs" - images that CLIP perceives as similar despite their cle
ar visual differences. With these pairs we construct the Multimodal Visual Patte
rns (MMVP) benchmark. MMVP exposes areas where state-of-the-art systems includin
g GPT-4V struggle with straightforward questions across nine basic visual patter
ns often providing incorrect answers and hallucinated explanations. We further e
valuate various CLIP-based vision-and-language models and found a notable correl
ation between visual patterns that challenge CLIP models and those problematic f
or multimodal LLMs. As an initial effort to address these issues we propose a Mi

xture of Features (MoF) approach demonstrating that integrating vision self-supe rvised learning features with MLLMs can significantly enhance their visual grounding capabilities. Together our research suggests visual representation learning remains an open challenge and accurate visual grounding is crucial for future successful multimodal systems.
****
Mean-Shift Feature Transformer
Takumi Kobayashi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6047-6056
Transformer models developed in NLP make a great impact on computer vision fields producing promising performance on various tasks. While multi-head attention a characteristic mechanism of the transformer attracts keen research interest such as for reducing computation cost we analyze the transformer model from a viewpoint of feature transformation based on a distribution of input feature tokens. The analysis inspires us to derive a novel transformation method from mean-shift update which is an effective gradient ascent to seek a local mode of distinctive representation on the token distribution. We also present an efficient projection approach to reduce parameter size of linear projections constituting the proposed multi-head feature transformation. In the experiments on ImageNet-1K dataset the proposed methods embedded into various network models exhibit favorable performance improvement in place of the transformer module.
****
Domain Separation Graph Neural Networks for Saliency Object Ranking
Zijian Wu, Jun Lu, Jing Han, Lianfa Bai, Yi Zhang, Zhuang Zhao, Siyang Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3964-3974
Saliency object ranking (SOR) has attracted significant attention recently. Previous methods usually failed to explicitly explore the saliency degree-related relationships between objects. In this paper we propose a novel Domain Separation Graph Neural Network (DSGNN) which starts with separately extracting the shape and texture cues from each object and builds an shape graph as well as a texture graph for all objects in the given image. Then we propose a Shape-Texture Graph Domain Separation (STGDS) module to separate the task-relevant and irrelevant information of target objects by explicitly modelling the relationship between each pair of objects in terms of their shapes and textures respectively. Furthermore a Cross Image Graph Domain Separation (CIGDS) module is introduced to explore the saliency degree subspace that is robust to different scenes aiming to create a unified representation for targets with the same saliency levels in different images. Importantly our DSGNN automatically learns a multi-dimensional feature to represent each graph edge allowing complex diverse and ranking-related relationships to be modelled. Experimental results show that our DSGNN achieved the new state-of-the-art performance on both ASSR and IRSR datasets with large improvements of 5.2% and 4.1% SA-SOR respectively. Our code is provided in https://github.com/Wu-ZJ/DSGNN.
****
Mind Marginal Non-Crack Regions: Clustering-Inspired Representation Learning for Crack Segmentation
Zhuangzhuang Chen, Zhuonan Lai, Jie Chen, Jianqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12698-12708
Crack segmentation datasets make great efforts to obtain the ground truth crack or non-crack labels as clearly as possible. However it can be observed that ambiguities are still inevitable when considering the marginal non-crack region due to low contrast and heterogeneous texture. To solve this problem we propose a novel clustering-inspired representation learning framework which contains a two-phase strategy for automatic crack segmentation. In the first phase a pre-process is proposed to localize the marginal non-crack region. Then we propose an ambiguity-aware segmentation loss (Aseg Loss) that enables crack segmentation models to capture ambiguities in the above regions via learning segmentation variance which allows us to further localize ambiguous regions. In the second phase to lea

rn the discriminative features of the above regions we propose a clustering-insp
ired loss (CI Loss) that alters the supervision learning of these regions into a
n unsupervised clustering manner. We demonstrate that the proposed method could
surpass the existing crack segmentation models on various datasets and our const
ructed CrackSeg5k dataset.
****
FISBe: A Real-World Benchmark Dataset for Instance Segmentation of Long-Range Th
in Filamentous Structures
Lisa Mais, Peter Hirsch, Claire Managan, Ramya Kandarpa, Josef Lorenz Rumberger,
 Annika Reinke, Lena Maier-Hein, Gudrun Ihrke, Dagmar Kainmueller; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 22249-22259
Instance segmentation of neurons in volumetric light microscopy images of nervou
s systems enables groundbreaking research in neuroscience by facilitating joint
functional and morphological analyses of neural circuits at cellular resolution.
 Yet said multi-neuron light microscopy data exhibits extremely challenging prop
erties for the task of instance segmentation: Individual neurons have long-rangi
ng thin filamentous and widely branching morphologies multiple neurons are tight
ly inter-weaved and partial volume effects uneven illumination and noise inheren
t to light microscopy severely impede local disentangling as well as long-range
tracing of individual neurons. These properties reflect a current key challenge
in machine learning research namely to effectively capture long-range dependenci
es in the data. While respective methodological research is buzzing to date meth
ods are typically benchmarked on synthetic datasets. To address this gap we rele
ase the FlyLight Instance Segmentation Benchmark (FISBe) dataset the first publi
cly available multi-neuron light microscopy dataset with pixel-wise annotations.
 In addition we define a set of instance segmentation metrics for benchmarking t
hat we designed to be meaningful with regard to downstream analyses. Lastly we p
rovide three baselines to kick off a competition that we envision to both advanc
e the field of machine learning regarding methodology for capturing long-range d
ata dependencies and facilitate scientific discovery in basic neuroscience.
****
RegionGPT: Towards Region Understanding Vision Language Model
Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou
Yu, Ping Luo, Sifei Liu; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 13796-13806
Vision language models (VLMs) have experienced rapid advancements through the in
tegration of large language models (LLMs) with image-text pairs yet they struggl
e with detailed regional visual understanding due to limited spatial awareness o
f the vision encoder and the use of coarse-grained training data that lacks deta
iled region-specific captions. To address this we introduce RegionGPT (short as
RGPT) a novel framework designed for complex region-level captioning and underst
anding. RGPT enhances the spatial awareness of regional representation with simp
le yet effective modifications to existing visual encoders in VLMs. We further i
mprove performance on tasks requiring a specific output scope by integrating tas
k-guided instruction prompts during both training and inference phases while mai
ntaining the model's versatility for general-purpose tasks. Additionally we deve
lop an automated region caption data generation pipeline enriching the training
set with detailed region-level captions. We demonstrate that a universal RGPT mo
del can be effectively applied and significantly enhancing performance across a
range of region-level tasks including but not limited to complex region descript
ions reasoning object classification and referring expressions comprehension.
****
LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding Reasoning
 and Planning
Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, J
iayuan Fan, Tao Chen; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2024, pp. 26428-26438
Recent progress in Large Multimodal Models (LMM) has opened up great possibiliti
es for various applications in the field of human-machine interactions. However

developing LMMs that can comprehend reason and plan in complex and diverse 3D environments remains a challenging topic especially considering the demand for understanding permutation-invariant point cloud representations of the 3D scene. Existing works seek help from multi-view images by projecting 2D features to 3D space which inevitably leads to huge computational overhead and performance degradation. In this paper we present LL3DA a Large Language 3D Assistant that takes point cloud as the direct input and responds to both text instructions and visual interactions. The additional visual interaction enables LMMs to better comprehend human interactions with the 3D environment and further remove the ambiguities within plain texts. Experiments show that LL3DA achieves remarkable results and surpasses various 3D vision-language models on both 3D Dense Captioning and 3D Question Answering.
****

4D Gaussian Splatting for Real-Time Dynamic Scene Rendering

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, Xinggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20310-20320

Representing and rendering dynamic scenes has been an important but challenging task. Especially to accurately model complex motions high efficiency is usually hard to guarantee. To achieve real-time dynamic scene rendering while also enjoying high training and storage efficiency we propose 4D Gaussian Splatting (4D-GS) as a holistic representation for dynamic scenes rather than applying 3D-GS for each individual frame. In 4D-GS a novel explicit representation containing both 3D Gaussians and 4D neural voxels is proposed. A decomposed neural voxel encoding algorithm inspired by HexPlane is proposed to efficiently build Gaussian features from 4D neural voxels and then a lightweight MLP is applied to predict Gaussian deformations at novel timestamps. Our 4D-GS method achieves real-time rendering under high resolutions 82 FPS at an 800*800 resolution on an RTX 3090 GPU while maintaining comparable or better quality than previous state-of-the-art methods. More demos and code are available at https://guanjunwu.github.io/4dgs.
****

RAM-Avatar: Real-time Photo-Realistic Avatar from Monocular Videos with Full-body Control

Xiang Deng, Zerong Zheng, Yuxiang Zhang, Jingxiang Sun, Chao Xu, Xiaodong Yang, Lizhen Wang, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1996-2007

This paper focuses on advancing the applicability of human avatar learning methods by proposing RAM-Avatar which learns a Real-time photo-realistic Avatar that supports full-body control from Monocular videos. To achieve this goal RAM-Avatar leverages two statistical templates responsible for modeling the facial expression and hand gesture variations while a sparsely computed dual attention module is introduced upon another body template to facilitate high-fidelity texture rendering for the torsos and limbs. Building on this foundation we deploy a lightweight yet powerful StyleUnet along with a temporal-aware discriminator to achieve real-time realistic rendering. To enable robust animation for out-of-distribution poses we propose a Motion Distribution Align module to compensate for the discrepancies between the training and testing motion distribution. Results and extensive experiments conducted in various experimental settings demonstrate the superiority of our proposed method and a real-time live system is proposed to further push research into applications. The training and testing code will be released for research purposes.
****

Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching

Xianqi Wang, Gangwei Xu, Hao Jia, Xin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19701-19710

Stereo matching methods based on iterative optimization like RAFT-Stereo and IGEV-Stereo have evolved into a cornerstone in the field of stereo matching. However these methods struggle to simultaneously capture high-frequency information in edges and low-frequency information in smooth regions due to the fixed receptive field. As a result they tend to lose details blur edges and produce false matc

hes in textureless areas. In this paper we propose Selective Recurrent Unit (SRU) a novel iterative update operator for stereo matching. The SRU module can adaptively fuse hidden disparity information at multiple frequencies for edge and smooth regions. To perform adaptive fusion we introduce a new Contextual Spatial Attention (CSA) module to generate attention maps as fusion weights. The SRU empowers the network to aggregate hidden disparity information across multiple frequencies mitigating the risk of vital hidden disparity information loss during iterative processes. To verify SRU's universality we apply it to representative iterative stereo matching methods collectively referred to as Selective-Stereo. Our Selective-Stereo ranks first on KITTI 2012 KITTI 2015 ETH3D and Middlebury leaderboards among all published methods. Code is available at https://github.com/Windsrain/Selective-Stereo.
****
PerAda: Parameter-Efficient Federated Learning Personalization with Generalization Guarantees
Chulin Xie, De-An Huang, Wenda Chu, Daguang Xu, Chaowei Xiao, Bo Li, Anima Anandkumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23838-23848
Personalized Federated Learning (pFL) has emerged as a promising solution to tackle data heterogeneity across clients in FL. However existing pFL methods either (1) introduce high computation and communication costs or (2) overfit to local data which can be limited in scope and vulnerable to evolved test samples with natural distribution shifts. In this paper we propose PerAda a parameter-efficient pFL framework that reduces communication and computational costs and exhibits superior generalization performance especially under test-time distribution shifts. PerAda reduces the costs by leveraging the power of pretrained models and only updates and communicates a small number of additional parameters from adapters. PerAda achieves high generalization by regularizing each client's personalized adapter with a global adapter while the global adapter uses knowledge distillation to aggregate generalized information from all clients. Theoretically we provide generalization bounds of PerAda and we prove its convergence to stationary points under non-convex settings. Empirically PerAda demonstrates higher personalized performance (+4.85% on CheXpert) and enables better out-of-distribution generalization (+5.23% on CIFAR-10-C) on different datasets across natural and medical domains compared with baselines while only updating 12.6% of parameters per model. Our code is available at https://github.com/NVlabs/PerAda.
****
MAFA: Managing False Negatives for Vision-Language Pre-training
Jaeseok Byun, Dohoon Kim, Taesup Moon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27314-27324
We consider a critical issue of false negatives in Vision- Language Pre-training (VLP) a challenge that arises from the inherent many-to-many correspondence of image-text pairs in large-scale web-crawled datasets. The presence of false negatives can impede achieving optimal performance and even lead to a significant performance drop. To address this challenge we propose MAFA (MAnaging FAlse negatives) which consists of two pivotal components building upon the recently developed GRouped mIni-baTch sampling (GRIT) strategy: 1) an efficient connection mining process that identifies and converts false negatives into positives and 2) label smoothing for the image-text contrastive (ITC) loss. Our comprehensive experiments verify the effectiveness of MAFA across multiple downstream tasks emphasizing the crucial role of addressing false negatives in VLP potentially even surpassing the importance of addressing false posi- tives. In addition the compatibility of MAFA with the recent BLIP-family model is also demonstrated. Code is available at https://github.com/jaeseokbyun/MAFA.
****
Video Prediction by Modeling Videos as Continuous Multi-Dimensional Processes
Gaurav Shrivastava, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7236-7245
Diffusion models have made significant strides in image generation mastering tasks such as unconditional image synthesis text-image translation and image-to-ima

ge conversions. However their capability falls short in the realm of video prediction mainly because they treat videos as a collection of independent images relying on external constraints such as temporal attention mechanisms to enforce temporal coherence. In our paper we introduce a novel model class that treats video as a continuous multi-dimensional process rather than a series of discrete frames. Through extensive experimentation we establish state-of-the-art performance in video prediction validated on benchmark datasets including KTH BAIR Human3.6M and UCF101.
****
PICTURE: PhotorealistIC virtual Try-on from UnconstRained dEsigns
Shuliang Ning, Duomin Wang, Yipeng Qin, Zirong Jin, Baoyuan Wang, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6976-6985
In this paper we propose a novel virtual try-on from unconstrained designs (ucVTON) task to enable photorealistic synthesis of personalized composite clothing on input human images. Unlike prior arts constrained by specific input types our method allows flexible specification of style (text or image) and texture (full garment cropped sections or texture patches) conditions. To address the entanglement challenge when using full garment images as conditions we develop a two-stage pipeline with explicit disentanglement of style and texture. In the first stage we generate a human parsing map reflecting the desired style conditioned on the input. In the second stage we composite textures onto the parsing map areas based on the texture input. To represent complex and non-stationary textures that have never been achieved in previous fashion editing works we first propose extracting hierarchical and balanced CLIP features and applying position encoding in VTON. Experiments demonstrate superior synthesis quality and personalization enabled by our method. The flexible control over style and texture mixing brings virtual try-on to a new level of user experience for online shopping and fashion design.
****
InfLoRA: Interference-Free Low-Rank Adaptation for Continual Learning
Yan-Shuo Liang, Wu-Jun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23638-23647
Continual learning requires the model to learn multiple tasks sequentially. In continual learning the model should possess the ability to maintain its performance on old tasks (stability) and the ability to adapt to new tasks continuously (plasticity). Recently parameter-efficient fine-tuning (PEFT) which involves freezing a pre-trained model and injecting a small number of learnable parameters to adapt to downstream tasks has gained increasing popularity in continual learning. Although existing continual learning methods based on PEFT have demonstrated superior performance compared to those not based on PEFT most of them do not consider how to eliminate the interference of the new task on the old tasks which inhibits the model from making a good trade-off between stability and plasticity. In this work we propose a new PEFT method called interference-free low-rank adaptation (InfLoRA) for continual learning. InfLoRA injects a small number of parameters to reparameterize the pre-trained weights and shows that fine-tuning these injected parameters is equivalent to fine-tuning the pre-trained weights within a subspace. Furthermore InfLoRA designs this subspace to eliminate the interference of the new task on the old tasks making a good trade-off between stability and plasticity. Experimental results show that InfLoRA outperforms existing state-of-the-art continual learning methods on multiple datasets.
****
Towards Robust 3D Pose Transfer with Adversarial Learning
Haoyu Chen, Hao Tang, Ehsan Adeli, Guoying Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2295-2304
3D pose transfer that aims to transfer the desired pose to a target mesh is one of the most challenging 3D generation tasks. Previous attempts rely on well-defined parametric human models or skeletal joints as driving pose sources. However to obtain those clean pose sources cumbersome but necessary pre-processing pipelines are inevitable hindering implementations of the real-time applications. Thi

s work is driven by the intuition that the robustness of the model can be enhanced by introducing adversarial samples into the training leading to a more invulnerable model to the noisy inputs which even can be further extended to directly handling the real-world data like raw point clouds/scans without intermediate processing. Furthermore we propose a novel 3D pose Masked Autoencoder (3D-PoseMAE) a customized MAE that effectively learns 3D extrinsic presentations (i.e. pose). 3D-PoseMAE facilitates learning from the aspect of extrinsic attributes by simultaneously generating adversarial samples that perturb the model and learning the arbitrary raw noisy poses via a multi-scale masking strategy. Both qualitative and quantitative studies show that the transferred meshes given by our network result in much better quality. Besides we demonstrate the strong generalizability of our method on various poses different domains and even raw scans. Experimental results also show meaningful insights that the intermediate adversarial samples generated in the training can successfully attack the existing pose transfer models.
****
Error Detection in Egocentric Procedural Task Videos
Shih-Po Lee, Zijia Lu, Zekun Zhang, Minh Hoai, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18655-18666
We present a new egocentric procedural error dataset containing videos with various types of errors as well as normal videos and propose a new framework for procedural error detection using error-free training videos only. Our framework consists of an action segmentation model and a contrastive step prototype learning module to segment actions and learn useful features for error detection. Based on the observation that interactions between hands and objects often inform action and error understanding we propose to combine holistic frame features with relations features which we learn by building a graph using active object detection followed by a Graph Convolutional Network. To handle errors unseen during training we use our contrastive step prototype learning to learn multiple prototypes for each step capturing variations of error-free step executions. At inference time we use feature-prototype similarities for error detection. By experiments on three datasets we show that our proposed framework outperforms state-of-the-art video anomaly detection methods for error detection and provides smooth action and error predictions.
****
EAGLE: Eigen Aggregation Learning for Object-Centric Unsupervised Semantic Segmentation
Chanyoung Kim, Woojung Han, Dayun Ju, Seong Jae Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3523-3533
Semantic segmentation has innately relied on extensive pixel-level annotated data leading to the emergence of unsupervised methodologies. Among them leveraging self-supervised Vision Transformers for unsupervised semantic segmentation (USS) has been making steady progress with expressive deep features. Yet for semantically segmenting images with complex objects a predominant challenge remains: the lack of explicit object-level semantic encoding in patch-level features. This technical limitation often leads to inadequate segmentation of complex objects with diverse structures. To address this gap we present a novel approach EAGLE which emphasizes object-centric representation learning for unsupervised semantic segmentation. Specifically we introduce EiCue a spectral technique providing semantic and structural cues through an eigenbasis derived from the semantic similarity matrix of deep image features and color affinity from an image. Further by incorporating our object-centric contrastive loss with EiCue we guide our model to learn object-level representations with intra- and inter-image object-feature consistency thereby enhancing semantic accuracy. Extensive experiments on COCO-Stuff Cityscapes and Potsdam-3 datasets demonstrate the state-of-the-art USS results of EAGLE with accurate and consistent semantic segmentation across complex scenes.
****

AVID: Any-Length Video Inpainting with Diffusion Model
Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, Licheng Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7162-7172
Recent advances in diffusion models have successfully enabled text-guided image inpainting. While it seems straightforward to extend such editing capability into the video domain there have been fewer works regarding text-guided video inpainting. Given a video a masked region at its initial frame and an editing prompt it requires a model to do infilling at each frame following the editing guidance while keeping the out-of-mask region intact. There are three main challenges in text-guided video inpainting: (i) temporal consistency of the edited video (ii) supporting different inpainting types at different structural fidelity levels and (iii) dealing with variable video length. To address these challenges we introduce Any-Length Video Inpainting with Diffusion Model dubbed as AVID. At its core our model is equipped with effective motion modules and adjustable structure guidance for fixed-length video inpainting. Building on top of that we propose a novel Temporal MultiDiffusion sampling pipeline with a middle-frame attention guidance mechanism facilitating the generation of videos with any desired duration. Our comprehensive experiments show our model can robustly deal with various inpainting types at different video duration ranges with high quality.
****
NoiseCollage: A Layout-Aware Text-to-Image Diffusion Model Based on Noise Cropping and Merging
Takahiro Shirakawa, Seiichi Uchida; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8921-8930
Layout-aware text-to-image generation is a task to generate multi-object images that reflect layout conditions in addition to text conditions. The current layout-aware text-to-image diffusion models still have several issues including mismatches between the text and layout conditions and quality degradation of generated images. This paper proposes a novel layout-aware text-to-image diffusion model called NoiseCollage to tackle these issues. During the denoising process NoiseCollage independently estimates noises for individual objects and then crops and merges them into a single noise. This operation helps avoid condition mismatches; in other words it can put the right objects in the right places. Qualitative and quantitative evaluations show that NoiseCollage outperforms several state-of-the-art models. These successful results indicate that the crop-and-merge operation of noises is a reasonable strategy to control image generation. We also show that NoiseCollage can be integrated with ControlNet to use edges sketches and pose skeletons as additional conditions. Experimental results show that this integration boosts the layout accuracy of ControlNet. The code is available at https://github.com/univ-esuty/noisecollage.
****
Uncertainty-Guided Never-Ending Learning to Drive
Lei Lai, Eshed Ohn-Bar, Sanjay Arora, John Seon Keun Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15088-15098
We present a highly scalable self-training framework for incrementally adapting vision-based end-to-end autonomous driving policies in a semi-supervised manner i.e. over a continual stream of incoming video data. To facilitate large-scale model training (e.g. open web or unlabeled data) we do not assume access to ground-truth labels and instead estimate pseudo-label policy targets for each video. Our framework comprises three key components: knowledge distillation a sample purification module and an exploration and knowledge retention mechanism. First given sequential image frames we pseudo-label the data and estimate uncertainty using an ensemble of inverse dynamics models. The uncertainty is used to select the most informative samples to add to an experience replay buffer. We specifically select high-uncertainty pseudo-labels to facilitate the exploration and learning of new and diverse driving skills. However in contrast to prior work in continual learning that assumes ground-truth labeled samples the uncertain pseudo-labels can introduce significant noise. Thus we also pair the exploration with a la

bel refinement module which makes use of consistency constraints to re-label the
 noisy exploratory samples and effectively learn from diverse data. Trained as a
 complete never-ending learning system we demonstrate state-of-the-art performan
ce on training from domain-changing data as well as millions of images from the
open web.
****
FakeInversion: Learning to Detect Images from Unseen Text-to-Image Models by Inv
erting Stable Diffusion
George Cazenavette, Avneesh Sud, Thomas Leung, Ben Usman; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10
759-10769
Due to the high potential for abuse of GenAI systems the task of detecting synth
etic images has recently become of great interest to the research community. Unf
ortunately existing image space detectors quickly become obsolete as new high-fi
delity text-to-image models are developed at blinding speed. In this work we pro
pose a new synthetic image detector that uses features obtained by inverting an
open-source pre-trained Stable Diffusion model. We show that these inversion fea
tures enable our detector to generalize well to unseen generators of high visual
 fidelity (e.g. DALL*E 3) even when the detector is trained only on lower fideli
ty fake images generated via Stable Diffusion. This detector achieves new state-
of-the-art across multiple training and evaluation setups. Moreover we introduce
 a new challenging evaluation protocol that uses reverse image search to mitigat
e stylistic and thematic biases in the detector evaluation. We show that the res
ulting evaluation scores align well with detectors' in-the-wild performance and
release these datasets as public benchmarks for future research.
****
PLGSLAM: Progressive Neural Scene Represenation with Local to Global Bundle Adju
stment
Tianchen Deng, Guole Shen, Tong Qin, Jianyu Wang, Wentao Zhao, Jingchuan Wang, D
anwei Wang, Weidong Chen; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2024, pp. 19657-19666
Neural implicit scene representations have recently shown encouraging results in
 dense visual SLAM. However existing methods produce low-quality scene reconstru
ction and low-accuracy localization performance when scaling up to large indoor
scenes and long sequences. These limitations are mainly due to their single glob
al radiance field with finite capacity which does not adapt to large scenarios.
Their end-to-end pose networks are also not robust enough with the growth of cum
ulative errors in large scenes. To this end we introduce PLGSLAM a neural visual
 SLAM system capable of high-fidelity surface reconstruction and robust camera t
racking in real-time. To handle large-scale indoor scenes PLGSLAM proposes a pro
gressive scene representation method which dynamically allocates new local scene
 representation trained with frames within a local sliding window. This allows u
s to scale up to larger indoor scenes and improves robustness (even under pose d
rifts). In local scene representation PLGSLAM utilizes tri-planes for local high
-frequency features with multi-layer perceptron (MLP) networks for the low-frequ
ency feature achieving smoothness and scene completion in unobserved areas. More
over we propose local-to-global bundle adjustment method with a global keyframe
database to address the increased pose drifts on long sequences. Experimental re
sults demonstrate that PLGSLAM achieves state-of-the-art scene reconstruction re
sults and tracking performance across various datasets and scenarios (both in sm
all and large-scale indoor environments).
****
Multi-Task Dense Prediction via Mixture of Low-Rank Experts
Yuqi Yang, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, Bo Li; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 27927-27937
Previous multi-task dense prediction methods based on the Mixture of Experts (Mo
E) have received great performance but they neglect the importance of explicitly
 modeling the global relations among all tasks. In this paper we present a novel
 decoder-focused method for multi-task dense prediction called Mixture-of-Low-Ra

nk-Experts (MLoRE). To model the global task relationships MLoRE adds a generic convolution path to the original MoE structure where each task feature can go through this path for explicit parameter sharing. Furthermore to control the parameters and computational cost brought by the increase in the number of experts we take inspiration from LoRA and propose to leverage the low-rank format of a vanilla convolution in the expert network. Since the low-rank experts have fewer parameters and can be dynamically parameterized into the generic convolution the parameters and computational cost do not change much with the increase of experts. Benefiting from this design we increase the number of experts and its reception field to enlarge the representation capacity facilitating multiple dense tasks learning in a unified network. Extensive experiments on the PASCAL-Context and NYUD-v2 benchmarks show that our MLoRE achieves superior performance compared to previous state-of-the-art methods on all metrics. Our code is available at https://github.com/YuqiYang213/MLoRE.
****
Binding Touch to Everything: Learning Unified Multimodal Tactile Representations
Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, Alex Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26340-26353
The ability to associate touch with other modalities has huge implications for humans and computational systems. However multimodal learning with touch remains challenging due to the expensive data collection process and non-standardized sensor outputs. We introduce UniTouch a unified tactile model for vision-based touch sensors connected to multiple modalities including vision language and sound. We achieve this by aligning our UniTouch embeddings to pretrained image embeddings already associated with a variety of other modalities. We further propose learnable sensor-specific tokens allowing the model to learn from a set of heterogeneous tactile sensors all at the same time. UniTouch is capable of conducting various touch sensing tasks in the zero-shot setting from robot grasping prediction to touch image question answering. To the best of our knowledge UniTouch is the first to demonstrate such capabilities.
****
Attribute-Guided Pedestrian Retrieval: Bridging Person Re-ID with Internal Attribute Variability
Yan Huang, Zhang Zhang, Qiang Wu, Yi Zhong, Liang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17689-17699
In various domains such as surveillance and smart retail pedestrian retrieval centering on person re-identification (Re-ID) plays a pivotal role. Existing Re-ID methodologies often overlook subtle internal attribute variations which are crucial for accurately identifying individuals with changing appearances. In response our paper introduces the Attribute-Guided Pedestrian Retrieval (AGPR) task focusing on integrating specified attributes with query images to refine retrieval results. Although there has been progress in attribute-driven image retrieval there remains a notable gap in effectively blending robust Re-ID models with intra-class attribute variations. To bridge this gap we present the Attribute-Guided Transformer-based Pedestrian Retrieval (ATPR) framework. ATPR adeptly merges global ID recognition with local attribute learning ensuring a cohesive linkage between the two. Furthermore to effectively handle the complexity of attribute interconnectivity ATPR organizes attributes into distinct groups and applies both inter-group correlation and intra-group decorrelation regularizations. Our extensive experiments on a newly established benchmark using the RAP dataset demonstrate the effectiveness of ATPR within the AGPR paradigm.
****
Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval
Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, Zhiqiang Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16551-16560
The increasing prevalence of video clips has sparked growing interest in text-vi

deo retrieval. Recent advances focus on establishing a joint embedding space for text and video relying on consistent embedding representations to compute similarity. However the text content in existing datasets is generally short and concise making it hard to fully describe the redundant semantics of a video. Correspondingly a single text embedding may be less expressive to capture the video embedding and empower the retrieval. In this study we propose a new stochastic text modeling method T-MASS i.e. text is modeled as a stochastic embedding to enrich text embedding with a flexible and resilient semantic range yielding a text mass. To be specific we introduce a similarity-aware radius module to adapt the scale of the text mass upon the given text-video pairs. Plus we design and develop a support text regularization to further control the text mass during the training. The inference pipeline is also tailored to fully exploit the text mass for accurate retrieval. Empirical evidence suggests that T-MASS not only effectively attracts relevant text-video pairs while distancing irrelevant ones but also enables the determination of precise text embeddings for relevant pairs. Our experimental results show a substantial improvement of T-MASS over baseline (3% 6.3% by R@1). Also T-MASS achieves state-of-the-art performance on five benchmark datasets including MSRVTT LSMDC DiDeMo VATEX and Charades.
****
Your Transferability Barrier is Fragile: Free-Lunch for Transferring the Non-Transferable Learning
Ziming Hong, Li Shen, Tongliang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28805-28815
Recently non-transferable learning (NTL) was proposed to restrict models' generalization toward the target domain(s) which serves as state-of-the-art solutions for intellectual property (IP) protection. However the robustness of the established "transferability barrier" for degrading the target domain performance has not been well studied. In this paper we first show that the generalization performance of NTL models is widely impaired on third-party domains (i.e. the unseen domain in the NTL training stage). We explore the impairment patterns and find that: due to the dominant generalization of non-transferable task NTL models tend to make target-domain-consistent predictions on third-party domains even though only a slight distribution shift from the third-party domain to the source domain. Motivated by these findings we uncover the potential risks of NTL by proposing a simple but effective method (dubbed as TransNTL) to recover the target domain performance with few source domain data. Specifically by performing a group of different perturbations on the few source domain data we obtain diverse third-party domains that evoke the same impairment patterns as the unavailable target domain. Then we fine-tune the NTL model under an impairment-repair self-distillation framework where the source-domain predictions are used to teach the model itself how to predict on third-party domains thus repairing the impaired generalization. Empirically experiments on standard NTL benchmarks show that the proposed TransNTL reaches up to 72% target-domain improvements by using only 10% source domain data. Finally we also explore a feasible defense method and empirically demonstrate its effectiveness.
****
Arbitrary Motion Style Transfer with Multi-condition Motion Latent Diffusion Model
Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, Hong Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 821-830
Computer animation's quest to bridge content and style has historically been a challenging venture with previous efforts often leaning toward one at the expense of the other. This paper tackles the inherent challenge of content-style duality ensuring a harmonious fusion where the core narrative of the content is both preserved and elevated through stylistic enhancements. We propose a novel Multi-condition Motion Latent Diffusion Model (MCM-LDM) for Arbitrary Motion Style Transfer (AMST). Our MCM-LDM significantly emphasizes preserving trajectories recognizing their fundamental role in defining the essence and fluidity of motion content. Our MCM-LDM's cornerstone lies in its ability first to disentangle and then

intricately weave together motion's tripartite components: motion trajectory motion content and motion style. The critical insight of MCM-LDM is to embed multiple conditions with distinct priorities. The content channel serves as the primary flow guiding the overall structure and movement while the trajectory and style channels act as auxiliary components and synchronize with the primary one dynamically. This mechanism ensures that multi-conditions can seamlessly integrate into the main flow enhancing the overall animation without overshadowing the core content. Empirical evaluations underscore the model's proficiency in achieving fluid and authentic motion style transfers setting a new benchmark in the realm of computer animation. The source code and model are available at https://github.com/XingliangJin/MCM-LDM.git.
****
Know Your Neighbors: Improving Single-View Reconstruction via Spatial Vision-Language Reasoning
Rui Li, Tobias Fischer, Mattia Segu, Marc Pollefeys, Luc Van Gool, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9848-9858
Recovering the 3D scene geometry from a single view is a fundamental yet ill-posed problem in computer vision. While classical depth estimation methods infer only a 2.5D scene representation limited to the image plane recent approaches based on radiance fields reconstruct a full 3D representation. However these methods still struggle with occluded regions since inferring geometry without visual observation requires (i) semantic knowledge of the surroundings and (ii) reasoning about spatial context. We propose KYN a novel method for single-view scene reconstruction that reasons about semantic and spatial context to predict each point's density. We introduce a vision-language modulation module to enrich point features with fine-grained semantic information. We aggregate point representations across the scene through a language-guided spatial attention mechanism to yield per-point density predictions aware of the 3D semantic context. We show that KYN improves 3D shape recovery compared to predicting density for each 3D point in isolation. We achieve state-of-the-art results in scene and object reconstruction on KITTI-360 and show improved zero-shot generalization compared to prior work. Project page: https://ruili3.github.io/kyn
****
Complementing Event Streams and RGB Frames for Hand Mesh Reconstruction
Jianping Jiang, Xinyu Zhou, Bingxuan Wang, Xiaoming Deng, Chao Xu, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24944-24954
Reliable hand mesh reconstruction (HMR) from commonly-used color and depth sensors is challenging especially under scenarios with varied illuminations and fast motions. Event camera is a highly promising alternative for its high dynamic range and dense temporal resolution properties but it lacks key texture appearance for hand mesh reconstruction. In this paper we propose EvRGBHand -- the first approach for 3D hand mesh reconstruction with an event camera and an RGB camera compensating for each other. By fusing two modalities of data across time space and information dimensionsEvRGBHand can tackle overexposure and motion blur issues in RGB-based HMR and foreground scarcity and background overflow issues in event-based HMR. We further propose EvRGBDegrader which allows our model to generalize effectively in challenging scenes even when trained solely on standard scenes thus reducing data acquisition costs. Experiments on real-world data demonstrate that EvRGBHand can effectively solve the challenging issues when using either type of camera alone via retaining the merits of both and shows the potential of generalization to outdoor scenes and another type of event camera. Our code models and dataset will be made public after acceptance.
****
Empowering Resampling Operation for Ultra-High-Definition Image Enhancement with Model-Aware Guidance
Wei Yu, Jie Huang, Bing Li, Kaiwen Zheng, Qi Zhu, Man Zhou, Feng Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25722-25731

Image enhancement algorithms have made remarkable advancements in recent years but directly applying them to Ultra-high-definition (UHD) images presents intractable computational overheads. Therefore previous straightforward solutions employ resampling techniques to reduce the resolution by adopting a "Downsampling-Enhancement-Upsampling" processing paradigm. However this paradigm disentangles the resampling operators and inner enhancement algorithms which results in the loss of information that is favored by the model further leading to sub-optimal outcomes. In this paper we propose a novel method of Learning Model-Aware Resampling (LMAR) which learns to customize resampling by extracting model-aware information from the UHD input image under the guidance of model knowledge. Specifically our method consists of two core designs namely compensatory kernel estimation and steganographic resampling. At the first stage we dynamically predict compensatory kernels tailored to the specific input and resampling scales. At the second stage the image-wise compensatory information is derived with the compensatory kernels and embedded into the rescaled input images. This promotes the representation of the newly derived downscaled inputs to be more consistent with the full-resolution UHD inputs as perceived by the model. Our LMAR enables model-aware and model-favored resampling while maintaining compatibility with existing resampling operators. Extensive experiments on multiple UHD image enhancement datasets and different backbones have shown consistent performance gains after correlating resizer and enhancer e.g. up to 1.2dB PSNR gain for x1.8 resampling scale on UHD-LOL4K. The code is available at \href https://github.com/YPatrickW/LMAR https://github.com/YPatrickW/LMAR .
****
ViT-CoMer: Vision Transformer with Convolutional Multi-scale Feature Interaction for Dense Predictions
Chunlong Xia, Xinliang Wang, Feng Lv, Xin Hao, Yifeng Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5493-5502
Although Vision Transformer (ViT) has achieved significant success in computer vision it does not perform well in dense prediction tasks due to the lack of inner-patch information interaction and the limited diversity of feature scale. Most existing studies are devoted to designing vision-specific transformers to solve the above problems which introduce additional pre-training costs. Therefore we present a plain pre-training-free and feature-enhanced ViT backbone with Convolutional Multi-scale feature interaction named ViT-CoMer which facilitates bidirectional interaction between CNN and transformer. Compared to the state-of-the-art ViT-CoMer has the following advantages: (1) We inject spatial pyramid multi-receptive field convolutional features into the ViT architecture which effectively alleviates the problems of limited local information interaction and single-feature representation in ViT. (2) We propose a simple and efficient CNN-Transformer bidirectional fusion interaction module that performs multi-scale fusion across hierarchical features which is beneficial for handling dense prediction tasks. (3) We evaluate the performance of ViT-CoMer across various dense prediction tasks different frameworks and multiple advanced pre-training. Notably our ViT-CoMer-L achieves 64.3% AP on COCO val2017 without extra training data and 62.1% mIoU on ADE20K val both of which are comparable to state-of-the-art methods. We hope ViT-CoMer can serve as a new backbone for dense prediction tasks to facilitate future research. The code will be released at https://github.com/Traffic-X/ViT-CoMer.
****
PromptCoT: Align Prompt Distribution via Adapted Chain-of-Thought
Junyi Yao, Yijiang Liu, Zhen Dong, Mingfei Guo, Helan Hu, Kurt Keutzer, Li Du, Daquan Zhou, Shanghang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7027-7037
Diffusion-based generative models have exhibited remarkable capability in the production of high-fidelity visual content such as images and videos. However their performance is significantly contingent upon the quality of textual inputs commonly referred to as "prompts". The process of traditional prompt engineering while effective necessitates empirical expertise and poses challenges for inexperi

enced users. In this paper we introduce PromptCoT an innovative enhancer that au tonomously refines prompts for users. PromptCoT is designed based on the observa tion that prompts which resemble the textual information of high-quality images in the training set often lead to superior generation performance. Therefore we fine-tune the pre-trained Large Language Models (LLM) using a curated text datas et that solely comprises descriptions of high-quality visual content. By doing s o the LLM can capture the distribution of high-quality training texts enabling i t to generate aligned continuations and revisions to boost the original texts. N onetheless one drawback of pre-trained LLMs is their tendency to generate extran eous or irrelevant information. We employ the Chain-of-Thought (CoT) mechanism t o improve the alignment between the original text prompts and their refined vers ions. CoT can extract and amalgamate crucial information from the aligned contin uation and revision enabling reasonable inferences based on the contextual cues to produce a more comprehensive and nuanced final output. Considering computatio nal efficiency instead of allocating a dedicated LLM for prompt enhancement to e ach individual model or dataset we integrate adapters that facilitate dataset-sp ecific adaptation leveraging a shared pre-trained LLM as the foundation for this process. With independent fine-tuning of these adapters we can adapt PromptCoT to new datasets while minimally increasing training costs and memory usage. We e valuate the effectiveness of PromptCoT by assessing its performance on widely-us ed latent diffusion models for image and video generation. The results demonstra te significant improvements in key performance metrics.
****
Hallucination Augmented Contrastive Learning for Multimodal Large Language Model
Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, Shikun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27036-27046
Multi-modal large language models (MLLMs) have been shown to efficiently integra te natural language with visual information to handle multi-modal tasks. However MLLMs still face a fundamental limitation of hallucinations where they tend to generate erroneous or fabricated information. In this paper we address hallucina tions in MLLMs from a novel perspective of representation learning. We first ana lyzed the representation distribution of textual and visual tokens in MLLM revea ling two important findings: 1) there is a significant gap between textual and v isual representations indicating unsatisfactory cross-modal representation align ment; 2) representations of texts that contain and do not contain hallucinations are entangled making it challenging to distinguish them. These two observations inspire us with a simple yet effective method to mitigate hallucinations. Speci fically we introduce contrastive learning into MLLMs and use text with hallucina tion as hard negative examples naturally bringing representations of non-halluci native text and visual samples closer while pushing way representations of non-h allucinating and hallucinative text. We evaluate our method quantitatively and q ualitatively showing its effectiveness in reducing hallucination occurrences and improving performance across multiple benchmarks. On the MMhal-Bench benchmark our method obtains a 34.66% /29.5% improvement over the baseline MiniGPT-4/LLaVA . Our code is available on https://github.com/X-PLUG/mPLUG-HalOwl/tree/main/hacl .
****
Preserving Fairness Generalization in Deepfake Detection
Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, Shu Hu; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16815 -16825
Although effective deepfake detection models have been developed in recent years recent studies have revealed that these models can result in unfair performance disparities among demographic groups such as race and gender. This can lead to particular groups facing unfair targeting or exclusion from detection potentiall y allowing misclassified deepfakes to manipulate public opinion and undermine tr ust in the model. The existing method for addressing this problem is providing a fair loss function. It shows good fairness performance for intra-domain evaluat ion but does not maintain fairness for cross-domain testing. This highlights the

significance of fairness generalization in the fight against deepfakes. In this work we propose the first method to address the fairness generalization problem in deepfake detection by simultaneously considering features loss and optimization aspects. Our method employs disentanglement learning to extract demographic and domain-agnostic forgery features fusing them to encourage fair learning across a flattened loss landscape. Extensive experiments on prominent deepfake datasets demonstrate our method's effectiveness surpassing state-of-the-art approaches in preserving fairness during cross-domain deepfake detection. The code is available at https://github.com/Purdue-M2/Fairness-Generalization.
****
Anomaly Score: Evaluating Generative Models and Individual Generated Images based on Complexity and Vulnerability
Jaehui Hwang, Junghyuk Lee, Jong-Seok Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8754-8763
With the advancement of generative models the assessment of generated images becomes increasingly more important. Previous methods measure distances between features of reference and generated images from trained vision models. In this paper we conduct an extensive investigation into the relationship between the representation space and input space around generated images. We first propose two measures related to the presence of unnatural elements within images: complexity which indicates how non-linear the representation space is and vulnerability which is related to how easily the extracted feature changes by adversarial input changes. Based on these we introduce a new metric to evaluating image-generative models called anomaly score (AS). Moreover we propose AS-i (anomaly score for individual images) that can effectively evaluate generated images individually. Experimental results demonstrate the validity of the proposed approach.
****
Structure-Aware Sparse-View X-ray 3D Reconstruction
Yuanhao Cai, Jiahao Wang, Alan Yuille, Zongwei Zhou, Angtian Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11174-11183
X-ray known for its ability to reveal internal structures of objects is expected to provide richer information for 3D reconstruction than visible light. Yet existing NeRF algorithms overlook this nature of X-ray leading to their limitations in capturing structural contents of imaged objects. In this paper we propose a framework Structure-Aware X-ray Neural Radiodensity Fields (SAX-NeRF) for sparse-view X-ray 3D reconstruction. Firstly we design a Line Segment-based Transformer (Lineformer) as the backbone of SAX-NeRF. Linefomer captures internal structures of objects in 3D space by modeling the dependencies within each line segment of an X-ray. Secondly we present a Masked Local-Global (MLG) ray sampling strategy to extract contextual and geometric information in 2D projection. Plus we collect a larger-scale dataset X3D covering wider X-ray applications. Experiments on X3D show that SAX-NeRF surpasses previous NeRF-based methods by 12.56 and 2.49 dB on novel view synthesis and CT reconstruction. https://github.com/caiyuanhao1998/SAX-NeRF
****
Dexterous Grasp Transformer
Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17933-17942
In this work we propose a novel discriminative framework for dexterous grasp generation named Dexterous Grasp TRansformer (DGTR) capable of predicting a diverse set of feasible grasp poses by processing the object point cloud with only one forward pass. We formulate dexterous grasp generation as a set prediction task and design a transformer-based grasping model for it. However we identify that this set prediction paradigm encounters several optimization challenges in the field of dexterous grasping and results in restricted performance. To address these issues we propose progressive strategies for both the training and testing phases. First the dynamic-static matching training (DSMT) strategy is presented to enhance the optimization stability during the training phase. Second we introduce

the adversarial-balanced test-time adaptation (AB-TTA) with a pair of adversarial losses to improve grasping quality during the testing phase. Experimental results on the DexGraspNet dataset demonstrate the capability of DGTR to predict dexterous grasp poses with both high quality and diversity. Notably while keeping high quality the diversity of grasp poses predicted by DGTR significantly outperforms previous works in multiple metrics without any data pre-processing. Codes are available at https://github.com/iSEE-Laboratory/DGTR.
****
Cooperation Does Matter: Exploring Multi-Order Bilateral Relations for Audio-Visual Segmentation
Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, Shiming Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27134-27143
Recently an audio-visual segmentation (AVS) task has been introduced aiming to group pixels with sounding objects within a given video. This task necessitates a first-ever audio-driven pixel-level understanding of the scene posing significant challenges. In this paper we propose an innovative audio-visual transformer framework termed COMBO an acronym for COoperation of Multi-order Bilateral relatiOns. For the first time our framework explores three types of bilateral entanglements within AVS: pixel entanglement modality entanglement and temporal entanglement. Regarding pixel entanglement we employ a Siam-Encoder Module (SEM) that leverages prior knowledge to generate more precise visual features from the foundational model. For modality entanglement we design a Bilateral-Fusion Module (BFM) enabling COMBO to align corresponding visual and auditory signals bi-directionally. As for temporal entanglement we introduce an innovative adaptive inter-frame consistency loss according to the inherent rules of temporal. Comprehensive experiments and ablation studies on AVSBench-object (84.7 mIoU on S4 59.2 mIou on MS3) and AVSBench-semantic (42.1 mIoU on AVSS) datasets demonstrate that COMBO surpasses previous state-of-the-art methods. Project page is available at https://yannqi.github.io/AVS-COMBO.
****
EgoThink: Evaluating First-Person Perspective Thinking Capability of Vision-Language Models
Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14291-14302
Vision-language models (VLMs) have recently shown promising results in traditional downstream tasks. Evaluation studies have emerged to assess their abilities with the majority focusing on the third-person perspective and only a few addressing specific tasks from the first-person perspective. However the capability of VLMs to "think" from a first-person perspective a crucial attribute for advancing autonomous agents and robotics remains largely unexplored. To bridge this research gap we introduce EgoThink a novel visual question-answering benchmark that encompasses six core capabilities with twelve detailed dimensions. The benchmark is constructed using selected clips from egocentric videos with manually annotated question-answer pairs containing first-person information. To comprehensively assess VLMs we evaluate twenty-one popular VLMs on EgoThink. Moreover given the open-ended format of the answers we use GPT-4 as the automatic judge to compute single-answer grading. Experimental results indicate that although GPT-4V leads in numerous dimensions all evaluated VLMs still possess considerable potential for improvement in first-person perspective tasks. Meanwhile enlarging the number of trainable parameters has the most significant impact on model performance on EgoThink. In conclusion EgoThink serves as a valuable addition to existing evaluation benchmarks for VLMs providing an indispensable resource for future research in the realm of embodied artificial intelligence and robotics.
****
Hearing Anything Anywhere
Mason Long Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11790-11799

Recent years have seen immense progress in 3D computer vision and computer graph
ics with emerging tools that can virtualize real-world 3D environments for numer
ous Mixed Reality (XR) applications. However alongside immersive visual experien
ces immersive auditory experiences are equally vital to our holistic perception
of an environment. In this paper we aim to reconstruct the spatial acoustic char
acteristics of an arbitrary environment given only a sparse set of (roughly 12)
room impulse response (RIR) recordings and a planar reconstruction of the scene
a setup that is easily achievable by ordinary users. To this end we introduce Di
ffRIR a differentiable RIR rendering framework with interpretable parametric mod
els of salient acoustic features of the scene including sound source directivity
 and surface reflectivity. This allows us to synthesize novel auditory experienc
es through the space with any source audio. To evaluate our method we collect a
dataset of RIR recordings and music in four diverse real environments. We show t
hat our model outperforms state-of-the-art baselines on rendering monaural and b
inaural RIRs and music at unseen locations and learns physically interpretable p
arameters characterizing acoustic properties of the sound source and surfaces in
 the scene.
****
PatchFusion: An End-to-End Tile-Based Framework for High-Resolution Monocular Me
tric Depth Estimation
Zhenyu Li, Shariq Farooq Bhat, Peter Wonka; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10016-10025
Single image depth estimation is a foundational task in computer vision and gene
rative modeling. However prevailing depth estimation models grapple with accommo
dating the increasing resolutions commonplace in today's consumer cameras and de
vices. Existing high-resolution strategies show promise but they often face limi
tations ranging from error propagation to the loss of high-frequency details. We
 present PatchFusion a novel tile-based framework with three key components to i
mprove the current state of the art: (1) A patch-wise fusion network that fuses
a globally-consistent coarse prediction with finer inconsistent tiled prediction
s via high-level feature guidance (2) A Global-to-Local (G2L) module that adds v
ital context to the fusion network discarding the need for patch selection heuri
stics and (3) A Consistency-Aware Training (CAT) and Inference (CAI) approach em
phasizing patch overlap consistency and thereby eradicating the necessity for po
st-processing. Experiments on UnrealStereo4K MVS-Synth and Middleburry 2014 demo
nstrate that our framework can generate high-resolution depth maps with intricat
e details. PatchFusion is independent of the base model for depth estimation. No
tably our framework built on top of SOTA ZoeDepth brings improvements for a tota
l of 17.3% and 29.4% in terms of the root mean squared error (RMSE) on UnrealSte
reo4K and MVS-Synth respectively.
****
GeneAvatar: Generic Expression-Aware Volumetric Head Avatar Editing from a Singl
e Image
Chong Bao, Yinda Zhang, Yuan Li, Xiyu Zhang, Bangbang Yang, Hujun Bao, Marc Poll
efeys, Guofeng Zhang, Zhaopeng Cui; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2024, pp. 8952-8963
Recently we have witnessed the explosive growth of various volumetric representa
tions in modeling animatable head avatars. However due to the diversity of frame
works there is no practical method to support high-level applications like 3D he
ad avatar editing across different representations. In this paper we propose a g
eneric avatar editing approach that can be universally applied to various 3DMM d
riving volumetric head avatars. To achieve this goal we design a novel expressio
n-aware modification generative model which enables lift 2D editing from a singl
e image to a consistent 3D modification field. To ensure the effectiveness of th
e generative modification process we develop several techniques including an exp
ression-dependent modification distillation scheme to draw knowledge from the la
rge-scale head avatar model and 2D facial texture editing tools implicit latent
space guidance to enhance model convergence and a segmentation-based loss reweig
ht strategy for fine-grained texture inversion. Extensive experiments demonstrat
e that our method delivers high-quality and consistent results across multiple e

xpression and viewpoints. Project page: https://zju3dv.github.io/ geneavatar/.
****
Improved Self-Training for Test-Time Adaptation
Jing Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2024, pp. 23701-23710
Test-time adaptation (TTA) is a technique to improve the performance of a pre-tr ained source model on a target distribution without using any labeled data. Howe ver existing self-trained TTA methods often face the challenges of unreliable ps eudo-labels and unstable model optimization. In this paper we propose an Improve d Self-Training (IST) approach which addresses these challenges by enhancing the pseudo-label quality and stabilizing the adaptation process. Specifically we us e a simple augmentation strategy to generate multiple views of each test sample and construct a graph structure to correct the pseudo-labels based on the simila rity of the latent features. Moreover we adopt a parameter moving average scheme to smooth the model updates and prevent catastrophic forgetting. Instead of usi ng a model with fixed label space we explore the adaptability of the foundation model CLIP to various downstream tasks at test time. Extensive experiments on va rious benchmarks show that IST can achieve significant and consistent improvemen ts over the existing TTA methods in classification detection and segmentation ta sks.
****
Learn to Rectify the Bias of CLIP for Unsupervised Semantic Segmentation
Jingyun Wang, Guoliang Kang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4102-4112
Recent works utilize CLIP to perform the challenging unsupervised semantic segme ntation task where only images without annotations are available. However we obs erve that when adopting CLIP to such a pixel-level understanding task unexpected bias occurs. Previous works don't explicitly model such bias which largely cons trains the segmentation performance. In this paper we propose to explicitly mode l and rectify the bias existing in CLIP to facilitate the unsupervised semantic segmentation. Specifically we design a learnable "Reference" prompt to encode cl ass-preference bias and project the positional embedding of vision transformer t o represent space-preference bias. Via a simple element-wise subtraction we rect ify the logits of CLIP classifier. Based on the rectified logits we generate a s egmentation mask via a Gumbel-Softmax operation. Then a contrastive loss between masked visual feature and the text features of different classes is imposed to facilitate the effective bias modeling. To further improve the segmentation we d istill the knowledge from the rectified CLIP to the advanced segmentation archit ecture via minimizing our designed mask-guided feature-guided and text-guided lo ss terms. Extensive experiments on standard benchmarks demonstrate that our meth od performs favorably against previous state-of-the-arts. The implementation is available at https://github.com/dogehhh/ReCLIP.
****
Unsupervised Feature Learning with Emergent Data-Driven Prototypicality
Yunhui Guo, Youren Zhang, Yubei Chen, Stella X. Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23199-23 208
Given a set of images our goal is to map each image to a point in a feature spac e such that not only point proximity indicates visual similarity but where it is located directly encodes how prototypical the image is according to the dataset . Our key insight is to perform unsupervised feature learning in hyperbolic inst ead of Euclidean space where the distance between points still reflects image si milarity yet we gain additional capacity for representing prototypicality with t he location of the point: The closer it is to the origin the more prototypical i t is. The latter property is simply emergent from optimizing the metric learning objective: The image similar to many training instances is best placed at the c enter of corresponding points in Euclidean space but closer to the origin in hyp erbolic space. We propose an unsupervised feature learning algorithm in Hyperbol ic space with sphere pACKing. HACK first generates uniformly packed particles in the Poincar'e ball of hyperbolic space and then assigns each image uniquely to

a particle. With our feature mapper simply trained to spread out training instances in hyperbolic space we observe that images move closer to the origin with congealing - a warping process that aligns all the images and makes them appear more common and similar to each other validating our idea of unsupervised prototypicality discovery. We demonstrate that our data-driven prototypicality provides an easy and superior unsupervised instance selection to reduce sample complexity increase model generalization with atypical instances and robustness with typical ones.
****
Unlocking Pre-trained Image Backbones for Semantic Image Synthesis
Tariq Berrada Ifriqi, Jakob Verbeek, Camille Couprie, Karteek Alahari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7840-7849
Semantic image synthesis i.e. generating images from user-provided semantic label maps is an important conditional image generation task as it allows to control both the content as well as the spatial layout of generated images. Although diffusion models have pushed the state of the art in generative image modeling the iterative nature of their inference process makes them computationally demanding. Other approaches such as GANs are more efficient as they only need a single feed-forward pass for generation but the image quality tends to suffer when modeling large and diverse datasets. In this work we propose a new class of GAN discriminators for semantic image synthesis that generates highly realistic images by exploiting feature backbones pre-trained for tasks such as image classification. We also introduce a new generator architecture with better context modeling and using cross-attention to inject noise into latent variables leading to more diverse generated images. Our model which we dub DP-SIMS achieves state-of-the-art results in terms of image quality and consistency with the input label maps on ADE-20K COCO-Stuff and Cityscapes surpassing recent diffusion models while requiring two orders of magnitude less compute for inference.
****
Retrieval-Augmented Egocentric Video Captioning
Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, Weidi Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13525-13536
Understanding human actions from videos of first-person view poses significant challenges. Most prior approaches explore representation learning on egocentric videos only while overlooking the potential benefit of exploiting existing large-scale third-person videos. In this paper (1) we develop EgoInstructor a retrieval-augmented multimodal captioning model that automatically retrieves semantically relevant third-person instructional videos to enhance the video captioning of egocentric videos (2) for training the cross-view retrieval module we devise an automatic pipeline to discover ego-exo video pairs from distinct large-scale egocentric and exocentric datasets (3) we train the cross-view retrieval module with a novel EgoExoNCE loss that pulls egocentric and exocentric video features closer by aligning them to shared text features that describe similar actions (4) through extensive experiments our cross-view retrieval module demonstrates superior performance across seven benchmarks. Regarding egocentric video captioning EgoInstructor exhibits significant improvements by leveraging third-person videos as references.
****
SkillDiffuser: Interpretable Hierarchical Planning via Skill Abstractions in Diffusion-Based Task Execution
Zhixuan Liang, Yao Mu, Hengbo Ma, Masayoshi Tomizuka, Mingyu Ding, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16467-16476
Diffusion models have demonstrated strong potential for robotic trajectory planning. However generating coherent trajectories from high-level instructions remains challenging especially for long-range composition tasks requiring multiple sequential skills. We propose SkillDiffuser an end-to-end hierarchical planning framework integrating interpretable skill learning with conditional diffusion plan

ning to address this problem. At the higher level the skill abstraction module learns discrete human-understandable skill representations from visual observations and language instructions. These learned skill embeddings are then used to condition the diffusion model to generate customized latent trajectories aligned with the skills. This allows generating diverse state trajectories that adhere to the learnable skills. By integrating skill learning with conditional trajectory generation SkillDiffuser produces coherent behavior following abstract instructions across diverse tasks. Experiments on multi-task robotic manipulation benchmarks like Meta-World and LOReL demonstrate state-of-the-art performance and human-interpretable skill representations from SkillDiffuser. More visualization results and information could be found on https://skilldiffuser.github.io/.
****
Improving Generalized Zero-Shot Learning by Exploring the Diverse Semantics from External Class Names
Yapeng Li, Yong Luo, Zengmao Wang, Bo Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23344-23353
Generalized Zero-Shot Learning (GZSL) methods often assume that the unseen classes are similar to seen classes and thus perform poor when unseen classes are dissimilar to seen classes. Although some existing GZSL approaches can alleviate this issue by leveraging additional semantic information from test unseen classes their generalization ability to dissimilar unseen classes is still unsatisfactory. This motivates us to study GZSL in the more practical setting where unseen classes can be either similar or dissimilar to seen classes. In this paper we propose a simple yet effective GZSL framework by exploring diverse semantics from external class names (DSECN) which is simultaneously robust on the similar and dissimilar unseen classes. This is achieved by introducing diverse semantics from external class names and aligning the introduced semantics to visual space using the classification head of pre-trained network. Furthermore we show that the design idea of DSECN can easily be integrate into other advanced GZSL approaches such as the generative-based ones and enhance their robustness for dissimilar unseen classes. Extensive experiments in the practical setting including both similar and dissimilar unseen classes show that our method significantly outperforms the state-of-the-art approaches on all datasets and can be trained very efficiently.
****
TeMO: Towards Text-Driven 3D Stylization for Multi-Object Meshes
Xuying Zhang, Bo-Wen Yin, Yuming Chen, Zheng Lin, Yunheng Li, Qibin Hou, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19531-19540
Recent progress in the text-driven 3D stylization of a single object has been considerably promoted by CLIP-based methods. However the stylization of multi-object 3D scenes is still impeded in that the image-text pairs used for pre-training CLIP mostly consist of an object. Meanwhile the local details of multiple objects may be susceptible to omission due to the existing supervision manner primarily relying on coarse-grained contrast of image-text pairs. To overcome these challenges we present a novel framework dubbed TeMO to parse multi-object 3D scenes and edit their styles under the contrast supervision at multiple levels. We first propose a Decoupled Graph Attention (DGA) module to distinguishably reinforce the features of 3D surface points. Particularly a cross-modal graph is constructed to align the object points accurately and noun phrases decoupled from the 3D mesh and textual description. Then we develop a Cross-Grained Contrast (CGC) supervision system where a fine-grained loss between the words in the textual description and the randomly rendered images are constructed to complement the coarse-grained loss. Extensive experiments show that our method can synthesize high-quality stylized content and outperform the existing methods over a wide range of multi-object 3D meshes.
****
TE-TAD: Towards Full End-to-End Temporal Action Detection via Time-Aligned Coordinate Expression
Ho-Joong Kim, Jung-Ho Hong, Heejo Kong, Seong-Whan Lee; Proceedings of the IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18837-18846
In this paper we investigate that the normalized coordinate expression is a key factor as reliance on hand-crafted components in query-based detectors for temporal action detection (TAD). Despite significant advancements towards an end-to-end framework in object detection query-based detectors have been limited in achieving full end-to-end modeling in TAD. To address this issue we propose TE-TAD a full end-to-end temporal action detection transformer that integrates time-aligned coordinate expression. We reformulate coordinate expression utilizing actual timeline values ensuring length-invariant representations from the extremely diverse video duration environment. Furthermore our proposed adaptive query selection dynamically adjusts the number of queries based on video length providing a suitable solution for varying video durations compared to a fixed query set. Our approach not only simplifies the TAD process by eliminating the need for hand-crafted components but also significantly improves the performance of query-based detectors. Our TE-TAD outperforms the previous query-based detectors and achieves competitive performance compared to state-of-the-art methods on popular benchmark datasets. Code is available at: https://github.com/Dotori-HJ/TE-TAD.
****
GSNeRF: Generalizable Semantic Neural Radiance Fields with Enhanced 3D Scene Understanding
Zi-Ting Chou, Sheng-Yu Huang, I-Jieh Liu, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20806-20815
Utilizing multi-view inputs to synthesize novel-view images Neural Radiance Fields (NeRF) have emerged as a popular research topic in 3D vision. In this work we introduce a Generalizable Semantic Neural Radiance Field (GSNeRF) which uniquely takes image semantics into the synthesis process so that both novel view images and the associated semantic maps can be produced for unseen scenes. Our GSNeRF is composed of two stages: Semantic Geo-Reasoning and Depth-Guided Visual rendering. The former is able to observe multi-view image inputs to extract semantic and geometry features from a scene. Guided by the resulting image geometry information the latter performs both image and semantic rendering with improved performances. Our experiments not only confirm that GSNeRF performs favorably against prior works on both novel-view image and semantic segmentation synthesis but the effectiveness of our sampling strategy for visual rendering is further verified.
****
Alpha Invariance: On Inverse Scaling Between Distance and Volume Density in Neural Radiance Fields
Joshua Ahn, Haochen Wang, Raymond A. Yeh, Greg Shakhnarovich; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20396-20405
Scale-ambiguity in 3D scene dimensions leads to magnitude-ambiguity of volumetric densities in neural radiance fields i.e. the densities double when scene size is halved and vice versa. We call this property alpha invariance. For NeRFs to better maintain alpha invariance we recommend 1) parameterizing both distance and volume densities in log space and 2) a discretization-agnostic initialization strategy to guarantee high ray transmittance. We revisit a few popular radiance field models and find that these systems use various heuristics to deal with issues arising from scene scaling. We test their behaviors and show our recipe to be more robust.
****
TexTile: A Differentiable Metric for Texture Tileability
Carlos Rodriguez-Pardo, Dan Casas, Elena Garces, Jorge Lopez-Moreno; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4439-4449
We introduce TexTile a novel differentiable metric to quantify the degree upon which a texture image can be concatenated with itself without introducing repeating artifacts (i.e. the tileability). Existing methods for tileable texture synth

esis focus on general texture quality but lack explicit analysis of the intrinsi
c repeatability properties of a texture. In contrast our TexTile metric effectiv
ely evaluates the tileable properties of a texture opening the door to more info
rmed synthesis and analysis of tileable textures. Under the hood TexTile is form
ulated as a binary classifier carefully built from a large dataset of textures o
f different styles semantics regularities and human annotations.Key to our metho
d is a set of architectural modifications to baseline pre-train image classifier
s to overcome their shortcomings at measuring tileability along with a custom da
ta augmentation and training regime aimed at increasing robustness and accuracy.
 We demonstrate that TexTile can be plugged into different state-of-the-art text
ure synthesis methods including diffusion-based strategies and generate tileable
 textures while keeping or even improving the overall texture quality. Furthermo
re we show that TexTile can objectively evaluate any tileable texture synthesis
method whereas the current mix of existing metrics produces uncorrelated scores
which heavily hinders progress in the field.
****
D3T: Distinctive Dual-Domain Teacher Zigzagging Across RGB-Thermal Gap for Domai
n-Adaptive Object Detection
Dinh Phat Do, Taehoon Kim, Jaemin Na, Jiwon Kim, Keonho Lee, Kyunghwan Cho, Wonj
un Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2024, pp. 23313-23322
Domain adaptation for object detection typically entails transferring knowledge
from one visible domain to another visible domain. However there are limited stu
dies on adapting from the visible to the thermal domain because the domain gap b
etween the visible and thermal domains is much larger than expected and traditio
nal domain adaptation can not successfully facilitate learning in this situation
. To overcome this challenge we propose a Distinctive Dual-Domain Teacher (D3T)
framework that employs distinct training paradigms for each domain. Specifically
 we segregate the source and target training sets for building dual-teachers and
 successively deploy exponential moving average to the student model to individu
al teachers of each domain. The framework further incorporates a zigzag learning
 method between dual teachers facilitating a gradual transition from the visible
 to thermal domains during training. We validate the superiority of our method t
hrough newly designed experimental protocols with well-known thermal datasets i.
e. FLIR and KAIST. Source code is available at https://github.com/EdwardDo69/D3T
.
****
Positive-Unlabeled Learning by Latent Group-Aware Meta Disambiguation
Lin Long, Haobo Wang, Zhijie Jiang, Lei Feng, Chang Yao, Gang Chen, Junbo Zhao;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2024, pp. 23138-23147
Positive-Unlabeled (PU) learning aims to train a binary classifier using minimal
 positive data supplemented by a substantially larger pool of unlabeled data in
the specific absence of explicitly annotated negatives. Despite its straightforw
ard nature as a binary classification task the currently best-performing PU algo
rithms still largely lag behind the supervised counterpart. In this work we iden
tify that the primary bottleneck lies in the difficulty of deriving discriminati
ve representations under unreliable binary supervision with poor semantics which
 subsequently hinders the common label disambiguation procedures. To cope with t
his problem we propose a novel PU learning framework namely Latent Group-Aware M
eta Disambiguation (LaGAM) which incorporates a hierarchical contrastive learnin
g module to extract the underlying grouping semantics within PU data and produce
 compact representations. As a result LaGAM enables a more aggressive label disa
mbiguation strategy where we enhance the robustness of training by iteratively d
istilling the true labels of unlabeled data directly through meta-learning. Exte
nsive experiments show that LaGAM significantly outperforms the current state-of
-the-art methods by an average of 6.8% accuracy on common benchmarks approaching
 the supervised baseline. We also provide comprehensive ablations as well as vis
ualized analysis to verify the effectiveness of our LaGAM.
****

Improving Image Restoration through Removing Degradations in Textual Representations
Jingbo Lin, Zhilu Zhang, Yuxiang Wei, Dongwei Ren, Dongsheng Jiang, Qi Tian, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2866-2878
In this paper we introduce a new perspective for improving image restoration by removing degradation in the textual representations of a given degraded image. Intuitively restoration is much easier on text modality than image one. For example it can be easily conducted by removing degradation-related words while keeping the content-aware words. Hence we combine the advantages of images in detail description and ones of text in degradation removal to perform restoration. To address the cross-modal assistance we propose to map the degraded images into textual representations for removing the degradations and then convert the restored textual representations into a guidance image for assisting image restoration. In particular We ingeniously embed an image-to-text mapper and text restoration module into CLIP-equipped text-to-image models to generate the guidance. Then we adopt a simple coarse-to-fine approach to dynamically inject multi-scale information from guidance to image restoration networks. Extensive experiments are conducted on various image restoration tasks including deblurring dehazing deraining and denoising and all-in-one image restoration. The results showcase that our method outperforms state-of-the-art ones across all these tasks. The codes and models are available at https://github.com/mrluin/TextualDegRemoval.
****
ZONE: Zero-Shot Instruction-Guided Local Editing
Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xiuhui Liu, Jiaming Liu, Lin Li, Xu Tang, Yao Hu, Jianzhuang Liu, Baochang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6254-6263
Recent advances in vision-language models like Stable Diffusion have shown remarkable power in creative image synthesis and editing.However most existing text-to-image editing methods encounter two obstacles: First the text prompt needs to be carefully crafted to achieve good results which is not intuitive or user-friendly. Second they are insensitive to local edits and can irreversibly affect non-edited regions leaving obvious editing traces. To tackle these problems we propose a Zero-shot instructiON-guided local image Editing approach termed ZONE. We first convert the editing intent from the user-provided instruction (e.g. "make his tie blue") into specific image editing regions through InstructPix2Pix. We then propose a Region-IoU scheme for precise image layer extraction from an off-the-shelf segment model. We further develop an edge smoother based on FFT for seamless blending between the layer and the image.Our method allows for arbitrary manipulation of a specific region with a single instruction while preserving the rest. Extensive experiments demonstrate that our ZONE achieves remarkable local editing results and user-friendliness outperforming state-of-the-art methods. Code is available at https://github.com/lsl001006/ZONE.
****
U-VAP: User-specified Visual Appearance Personalization via Decoupled Self Augmentation
You Wu, Kean Liu, Xiaoyue Mi, Fan Tang, Juan Cao, Jintao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9482-9491
Concept personalization methods enable large text-to-image models to learn specific subjects (e.g. objects/poses/3D models) and synthesize renditions in new contexts. Given that the image references are highly biased towards visual attributes state-of-the-art personalization models tend to overfit the whole subject and cannot disentangle visual characteristics in pixel space. In this study we proposed a more challenging setting namely fine-grained visual appearance personalization. Different from existing methods we allow users to provide a sentence describing the desired attributes. A novel decoupled self-augmentation strategy is proposed to generate target-related and non-target samples to learn user-specified visual attributes.These augmented data allow for refining the model's understa

nding of the target attribute while mitigating the impact of unrelated attribute s. At the inference stage adjustments are conducted on semantic space through th e learned target and non-target embeddings to further enhance the disentanglemen t of target attributes. Extensive experiments on various kinds of visual attribu tes with SOTA personalization methods shows the ability of the proposed method t o mimic target visual appearance in novel contexts thus improving the controllab ility and flexibility of personalization.
****
PointBeV: A Sparse Approach for BeV Predictions
Loick Chambon, Eloi Zablocki, Mickaël Chen, Florent Bartoccioni, Patrick Pérez, Matthieu Cord; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2024, pp. 15195-15204
Bird's-eye View (BeV) representations have emerged as the de-facto shared space in driving applications offering a unified space for sensor data fusion and supp orting various downstream tasks. However conventional models use grids with fixe d resolution and range and face computational inefficiencies due to the uniform allocation of resources across all cells. To address this we propose PointBeV a novel sparse BeV segmentation model operating on sparse BeV cells instead of den se grids. This approach offers precise control over memory usage enabling the us e of long temporal contexts and accommodating memory-constrained platforms. Poin tBeV employs an efficient two-pass strategy for training enabling focused comput ation on regions of interest. At inference time it can be used with various memo ry/performance trade-offs and flexibly adjusts to new specific use cases. PointB eV achieves state-of-the-art results on the nuScenes dataset for vehicle pedestr ian and lane segmentation showcasing superior performance in static and temporal settings despite being trained solely with sparse signals. We release our code with two new efficient modules used in the architecture: Sparse Feature Pulling designed for the effective extraction of features from images to BeV and Submani fold Attention which enables efficient temporal modeling. The code is available at https://github.com/valeoai/PointBeV.
****
From-Ground-To-Objects: Coarse-to-Fine Self-supervised Monocular Depth Estimatio n of Dynamic Objects with Ground Contact Prior
Jaeho Moon, Juan Luis Gonzalez Bello, Byeongjun Kwon, Munchurl Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 24, pp. 10519-10529
Self-supervised monocular depth estimation (DE) is an approach to learning depth without costly depth ground truths. However it often struggles with moving obje cts that violate the static scene assumption during training. To address this is sue we introduce a coarse-to-fine training strategy leveraging the ground contac ting prior based on the observation that most moving objects in outdoor scenes c ontact the ground. In the coarse training stage we exclude the objects in dynami c classes from the reprojection loss calculation to avoid inaccurate depth learn ing. To provide precise supervision on the depth of the objects we present a nov el Ground-contacting-prior Disparity Smoothness Loss (GDS-Loss) that encourages a DE network to align the depth of the objects with their ground-contacting poin ts. Subsequently in the fine training stage we refine the DE network to learn th e detailed depth of the objects from the reprojection loss while ensuring accura te DE on the moving object regions by employing our regularization loss with a c ost-volume-based weighting factor. Our overall coarse-to-fine training strategy can easily be integrated with existing DE methods without any modifications sign ificantly enhancing DE performance on challenging Cityscapes and KITTI datasets especially in the moving object regions.
****
Linguistic-Aware Patch Slimming Framework for Fine-grained Cross-Modal Alignment
Zheren Fu, Lei Zhang, Hou Xia, Zhendong Mao; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26307-26316
Cross-modal alignment aims to build a bridge connecting vision and language. It is an important multi-modal task that efficiently learns the semantic similariti es between images and texts. Traditional fine-grained alignment methods heavily

rely on pre-trained object detectors to extract region features for subsequent r
egion-word alignment thereby incurring substantial computational costs for regio
n detection and error propagation issues for two-stage training. In this paper w
e focus on the mainstream vision transformer incorporating patch features for pa
tch-word alignment while addressing the resultant issue of visual patch redundan
cy and patch ambiguity for semantic alignment. We propose a novel Linguistic-Awa
re Patch Slimming (LAPS) framework for fine-grained alignment which explicitly i
dentifies redundant visual patches with language supervision and rectifies their
 semantic and spatial information to facilitate more effective and consistent pa
tch-word alignment. Extensive experiments on various evaluation benchmarks and m
odel backbones show LAPS outperforms the state-of-the-art fine-grained alignment
 methods by 5%-15% rSum. Our code is available at https://github.com/CrossmodalG
roup/LAPS
****

HHMR: Holistic Hand Mesh Recovery by Enhancing the Multimodal Controllability of
 Graph Diffusion Models
Mengcheng Li, Hongwen Zhang, Yuxiang Zhang, Ruizhi Shao, Tao Yu, Yebin Liu; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 645-654
Recent years have witnessed a trend of the deep integration of the generation an
d reconstruction paradigms. In this paper we extend the ability of controllable
generative models for a more comprehensive hand mesh recovery task: direct hand
mesh generation inpainting reconstruction and fitting in a single framework whic
h we name as Holistic Hand Mesh Recovery (HHMR). Our key observation is that dif
ferent kinds of hand mesh recovery tasks can be achieved by a single generative
model with strong multimodal controllability and in such a framework realizing d
ifferent tasks only requires giving different signals as conditions. To achieve
this goal we propose an all-in-one diffusion framework based on graph convolutio
n and attention mechanisms for holistic hand mesh recovery. In order to achieve
strong control generation capability while ensuring the decoupling of multimodal
 control signals we map different modalities to a share feature space and apply
cross-scale random masking in both modality and feature levels. In this way the
correlation between different modalities can be fully exploited during the learn
ing of hand priors. Furthermore we propose Condition-aligned Gradient Guidance t
o enhance the alignment of the generated model with the control signals which si
gnificantly improves the accuracy of the hand mesh reconstruction and fitting. E
xperiments show that our novel framework can realize multiple hand mesh recovery
 tasks simultaneously and outperform the existing methods in different tasks whi
ch provides more possibilities for subsequent downstream applications including
gesture recognition pose generation mesh editing and so on.
****

SRTube: Video-Language Pre-Training with Action-Centric Video Tube Features and
Semantic Role Labeling
Ju-Hee Lee, Je-Won Kang; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 13689-13699
In recent years large-scale video-language pre-training (VidLP) has received con
siderable attention for its effectiveness in relevant tasks. In this paper we pr
opose a novel action-centric VidLP framework that employs video tube features fo
r temporal modeling and language features based on semantic role labeling (SRL).
 Our video encoder generates multiple tube features along object trajectories id
entifying action-related regions within videos to overcome the limitations of ex
isting temporal attention mechanisms. Additionally our text encoder incorporates
 high-level action-related language knowledge previously underutilized in curren
t VidLP models. The SRL captures action-verbs and related semantics among object
s in sentences and enhances the ability to perform instance-level text matching
thus enriching the cross-modal (CM) alignment process. We also introduce two nov
el pre-training objectives and a self-supervision strategy to produce a more fai
thful CM representation. Experimental results demonstrate that our method outper
forms existing VidLP frameworks in various downstream tasks and datasets establi
shing our model a baseline in the modern VidLP framework.

****
Prompt Highlighter: Interactive Control for Multi-Modal LLMs
Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13215-13224
This study targets a critical aspect of multi-modal LLMs' (LLMs&VLMs) inference: explicit controllable text generation. Multi-modal LLMs empower multi-modality understanding with the capability of semantic generation yet bring less explainability and heavier reliance on prompt contents due to their autoregressive generative nature. While manipulating prompt formats could improve outputs designing specific and precise prompts per task can be challenging and ineffective. To tackle this issue we introduce a novel inference method Prompt Highlighter which enables users to highlight specific prompt spans to interactively control the focus during generation. Motivated by the classifier-free diffusion guidance we form regular and unconditional context pairs based on highlighted tokens demonstrating that the autoregressive generation in models can be guided in a classifier-free way. Notably we find that during inference guiding the models with highlighted tokens through the attention weights leads to more desired outputs. Our approach is compatible with current LLMs and VLMs achieving impressive customized generation results without training. Experiments confirm its effectiveness in focusing on input contexts and generating reliable content. Without tuning on LLaVA-v1.5 our method secured 70.7 in the MMBench test and 1552.5 in MME-perception.
****
Domain-Rectifying Adapter for Cross-Domain Few-Shot Segmentation
Jiapeng Su, Qi Fan, Wenjie Pei, Guangming Lu, Fanglin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24036-24045
Few-shot semantic segmentation (FSS) has achieved great success on segmenting objects of novel classes supported by only a few annotated samples. However existing FSS methods often underperform in the presence of domain shifts especially when encountering new domain styles that are unseen during training. It is suboptimal to directly adapt or generalize the entire model to new domains in the few-shot scenario. Instead our key idea is to adapt a small adapter for rectifying diverse target domain styles to the source domain. Consequently the rectified target domain features can fittingly benefit from the well-optimized source domain segmentation model which is intently trained on sufficient source domain data. Training domain-rectifying adapter requires sufficiently diverse target domains. We thus propose a novel local-global style perturbation method to simulate diverse potential target domains by perturbating the feature channel statistics of the individual images and collective statistics of the entire source domain respectively. Additionally we propose a cyclic domain alignment module to facilitate the adapter effectively rectifying domains using a reverse domain rectification supervision. The adapter is trained to rectify the image features from diverse synthesized target domains to align with the source domain. During testing on target domains we start by rectifying the image features and then conduct few-shot segmentation on the domain-rectified features. Extensive experiments demonstrate the effectiveness of our method achieving promising results on cross-domain few-shot semantic segmentation tasks. Our code is available at https://github.com/Matt-Su/DR-Adapter.
****
Robust Self-calibration of Focal Lengths from the Fundamental Matrix
Viktor Kocur, Daniel Kyselica, Zuzana Kukelova; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5220-5229
The problem of self-calibration of two cameras from a given fundamental matrix is one of the basic problems in geometric computer vision. Under the assumption of known principal points and square pixels the Bougnoux formula offers a means to compute the two unknown focal lengths. However in many practical situations the formula yields inaccurate results due to commonly occurring singularities. Moreover the estimates are sensitive to noise in the computed fundamental matrix and to the assumed positions of the principal points. In this paper we therefore p

ropose an efficient and robust iterative method to estimate the focal lengths al
ong with the principal points of the cameras given a fundamental matrix and prio
rs for the estimated camera intrinsics. In addition we study a computationally e
fficient check of models generated within RANSAC that improves the accuracy of t
he estimated models while reducing the total computational time. Extensive exper
iments on real and synthetic data show that our iterative method brings signific
ant improvements in terms of the accuracy of the estimated focal lengths over th
e Bougnoux formula and other state-of-the-art methods even when relying on inacc
urate priors. The code for the methods and experiments is available at https://g
ithub.com/kocurvik/robust_self_calibration
****
Continual Learning for Motion Prediction Model via Meta-Representation Learning
and Optimal Memory Buffer Retention Strategy
DaeJun Kang, Dongsuk Kum, Sanmin Kim; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15438-15448
Embodied AI such as autonomous vehicles suffers from insufficient long-tailed da
ta because it must be obtained from the physical world. In fact data must be con
tinuously obtained in a series of small batches and the model must also be conti
nuously trained to achieve generalizability and scalability by improving the bia
sed data distribution. This paper addresses the training cost and catastrophic f
orgetting problems when continuously updating models to adapt to incoming small
batches from various environments for real-world motion prediction in autonomous
 driving. To this end we propose a novel continual motion prediction (CMP) learn
ing framework based on sparse meta-representation learning and an optimal memory
 buffer retention strategy. In meta-representation learning a model explicitly l
earns a sparse representation of each driving environment from road geometry to
vehicle states by training to reduce catastrophic forgetting based on an augment
ed modulation network with sparsity regularization. Also in the adaptation phase
 We develop an Optimal Memory Buffer Retention strategy that smartly preserves d
iverse samples by focusing on representation similarity. This approach handles t
he nuanced task distribution shifts characteristic of motion prediction datasets
 ensuring our model stays responsive to evolving input variations without requir
ing extensive resources. The experiment results demonstrate that the proposed me
thod shows superior adaptation performance to the conventional continual learnin
g approach which is developed using a synthetic dataset for the continual learni
ng problem.
****
PartDistill: 3D Shape Part Segmentation by Vision-Language Model Distillation
Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, Yen-Yu Lin; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2024, pp. 3470-3479
This paper proposes a cross-modal distillation framework PartDistill which trans
fers 2D knowledge from vision-language models (VLMs) to facilitate 3D shape part
 segmentation. PartDistill addresses three major challenges in this task: the la
ck of 3D segmentation in invisible or undetected regions in the 2D projections i
nconsistent 2D predictions by VLMs and the lack of knowledge accumulation across
 different 3D shapes. PartDistill consists of a teacher network that uses a VLM
to make 2D predictions and a student network that learns from the 2D predictions
 while extracting geometrical features from multiple 3D shapes to carry out 3D p
art segmentation. A bi-directional distillation including forward and backward d
istillations is carried out within the framework where the former forward distil
ls the 2D predictions to the student network and the latter improves the quality
 of the 2D predictions which subsequently enhances the final 3D segmentation. Mo
reover PartDistill can exploit generative models that facilitate effortless 3D s
hape creation for generating knowledge sources to be distilled. Through extensiv
e experiments PartDistill boosts the existing methods with substantial margins o
n widely used ShapeNetPart and PartNetE datasets by more than 15% and 12% higher
 mIoU scores respectively. The code for this work is available at https://github
.com/ardianumam/PartDistill.
****

CPP-Net: Embracing Multi-Scale Feature Fusion into Deep Unfolding CP-PPA Network for Compressive Sensing

Zhen Guo, Hongping Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25086-25095

In the domain of compressive sensing (CS) deep unfolding networks (DUNs) have garnered attention for their good performance and certain degree of interpretability rooted in CS domain achieved by marrying traditional optimization solvers with deep networks. However current DUNs are ill-suited for the intricate task of capturing fine-grained image details leading to perceptible distortions and blurriness in reconstructed images particularly at low CS ratios e.g. 0.10 and below. In this paper we propose CPP-Net a novel deep unfolding CS framework inspired by the primal-dual hybrid strategy of the Chambolle and Pock Proximal Point Algorithm (CP-PPA). First we derive three iteration submodules Xk Vk and Yk by incorporating customized deep learning modules to solve the sparse basis related proximal operator within CP-PPA. Second we design the Dual Path Fusion Block (DPFB) to adeptly extract and fuse multi-scale feature information enhancing sensitivity to feature information at different scales and improving detail reconstruction. Third we introduce the Iteration Fusion Strategy (IFS) to effectively weight the fusion of outputs from diverse reconstruction stages maximizing the utilization of feature information and mitigating the information loss during reconstruction stages. Extensive experiments demonstrate that CPP-Net effectively reduces distortion and blurriness while preserving richer image details outperforming current state-of-the-art methods. Codes are available at https://github.com/ICSResearch/CPP-Net.

****

EditGuard: Versatile Image Watermarking for Tamper Localization and Copyright Protection

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11964-11974

In the era of AI-generated content (AIGC) malicious tampering poses imminent threats to copyright integrity and information security. Current deep image watermarking while widely accepted for safeguarding visual content can only protect copyright and ensure traceability. They fall short in localizing increasingly realistic image tampering potentially leading to trust crises privacy violations and legal disputes. To solve this challenge we propose an innovative proactive forensics framework EditGuard to unify copyright protection and tamper-agnostic localization especially for AIGC-based editing methods. It can offer a meticulous embedding of imperceptible watermarks and precise decoding of tampered areas and copyright information. Leveraging our observed fragility and locality of image-into-image steganography the realization of EditGuard can be converted into a united image-bit steganography issue thus completely decoupling the training process from the tampering types. Extensive experiments verify that our EditGuard balances the tamper localization accuracy copyright recovery precision and generalizability to various AIGC-based tampering methods especially for image forgery that is difficult for the naked eye to detect.

****

3DGStream: On-the-Fly Training of 3D Gaussians for Efficient Streaming of Photo-Realistic Free-Viewpoint Videos

Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, Wei Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20675-20685

Constructing photo-realistic Free-Viewpoint Videos (FVVs) of dynamic scenes from multi-view videos remains a challenging endeavor. Despite the remarkable advancements achieved by current neural rendering techniques these methods generally require complete video sequences for offline training and are not capable of real-time rendering. To address these constraints we introduce 3DGStream a method designed for efficient FVV streaming of real-world dynamic scenes. Our method achieves fast on-the-fly per-frame reconstruction within 12 seconds and real-time rendering at 200 FPS. Specifically we utilize 3D Gaussians (3DGs) to represent the

scene. Instead of the naive approach of directly optimizing 3DGs per-frame we employ a compact Neural Transformation Cache (NTC) to model the translations and rotations of 3DGs markedly reducing the training time and storage required for each FVV frame. Furthermore we propose an adaptive 3DG addition strategy to handle emerging objects in dynamic scenes. Experiments demonstrate that 3DGStream achieves competitive performance in terms of rendering speed image quality training time and model storage when compared with state-of-the-art methods.
****
FairRAG: Fair Human Generation via Fair Retrieval Augmentation
Robik Shrestha, Yang Zou, Qiuyu Chen, Zhiheng Li, Yusheng Xie, Siqi Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11996-12005
Existing text-to-image generative models reflect or even amplify societal biases ingrained in their training data. This is especially concerning for human image generation where models are biased against certain demographic groups. Existing attempts to rectify this issue are hindered by the inherent limitations of the pre-trained models and fail to substantially improve demographic diversity. In this work we introduce Fair Retrieval Augmented Generation (FairRAG) a novel framework that conditions pre-trained generative models on reference images retrieved from an external image database to improve fairness in human generation. FairRAG enables conditioning through a lightweight linear module that projects reference images into the textual space. To enhance fairness FairRAG applies simple-yet-effective debiasing strategies providing images from diverse demographic groups during the generative process. Extensive experiments demonstrate that FairRAG outperforms existing methods in terms of demographic diversity image-text alignment and image fidelity while incurring minimal computational overhead during inference.
****
DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing
Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, Song Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8839-8849
Accurate and controllable image editing is a challenging task that has attracted significant attention recently. Notably DragGAN developed by Pan et al. (2023) is an interactive point-based image editing framework that achieves impressive editing results with pixel-level precision. However due to its reliance on generative adversarial networks (GANs) its generality is limited by the capacity of pretrained GAN models. In this work we extend this editing framework to diffusion models and propose a novel approach DragDiffusion. By harnessing large-scale pretrained diffusion models we greatly enhance the applicability of interactive point-based editing on both real and diffusion-generated images. Unlike other diffusion-based editing methods that provide guidance on diffusion latents of multiple time steps our approach achieves efficient yet accurate spatial control by optimizing the latent of only one time step. This novel design is motivated by our observations that UNet features at a specific time step provides sufficient semantic and geometric information to support the drag-based editing. Moreover we introduce two additional techniques namely identity-preserving fine-tuning and reference-latent-control to further preserve the identity of the original image. Lastly we present a challenging benchmark dataset called DragBench---the first benchmark to evaluate the performance of interactive point-based image editing methods. Experiments across a wide range of challenging cases (e.g. images with multiple objects diverse object categories various styles etc.) demonstrate the versatility and generality of DragDiffusion. Code and the DragBench dataset: https://github.com/Yujun-Shi/DragDiffusion.
****
FaceTalk: Audio-Driven Motion Diffusion for Neural Parametric Head Models
Shivangi Aneja, Justus Thies, Angela Dai, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21263-21273

We introduce FaceTalk a novel generative approach designed for synthesizing high
-fidelity 3D motion sequences of talking human heads from input audio signal. To
 capture the expressive detailed nature of human heads including hair ears and f
iner-scale eye movements we propose to couple speech signal with the latent spac
e of neural parametric head models to create high-fidelity temporally coherent m
otion sequences. We propose a new latent diffusion model for this task operating
 in the expression space of neural parametric head models to synthesize audio-dr
iven realistic head sequences. In the absence of a dataset with corresponding NP
HM expressions to audio we optimize for these correspondences to produce a datas
et of temporally-optimized NPHM expressions fit to audio-video recordings of peo
ple talking. To the best of our knowledge this is the first work to propose a ge
nerative approach for realistic and high-quality motion synthesis of volumetric
human heads representing a significant advancement in the field of audio-driven
3D animation. Notably our approach stands out in its ability to generate plausib
le motion sequences that can produce high-fidelity head animation coupled with t
he NPHM shape space. Our experimental results substantiate the effectiveness of
FaceTalk consistently achieving superior and visually natural motion encompassin
g diverse facial expressions and styles outperforming existing methods by 75% in
 perceptual user study evaluation
****
Mip-Splatting: Alias-free 3D Gaussian Splatting
Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, Andreas Geiger; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 19447-19456
Recently 3D Gaussian Splatting has demonstrated impressive novel view synthesis
results reaching high fidelity and efficiency. However strong artifacts can be o
bserved when changing the sampling rate e.g. by changing focal length or camera
distance. We find that the source for this phenomenon can be attributed to the l
ack of 3D frequency constraints and the usage of a 2D dilation filter. To addres
s this problem we introduce a 3D smoothing filter to constrains the size of the
3D Gaussian primitives based on the maximal sampling frequency induced by the in
put views. It eliminates high-frequency artifacts when zooming in. Moreover repl
acing 2D dilation with a 2D Mip filter which simulates a 2D box filter effective
ly mitigates aliasing and dilation issues. Our evaluation including scenarios su
ch a training on single-scale images and testing on multiple scales validates th
e effectiveness of our approach.
****
Learning Coupled Dictionaries from Unpaired Data for Image Super-Resolution
Longguang Wang, Juncheng Li, Yingqian Wang, Qingyong Hu, Yulan Guo; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 25712-25721
The difficulty of acquiring high-resolution (HR) and low-resolution (LR) image p
airs in real scenarios limits the performance of existing learning-based image s
uper-resolution (SR) methods in the real world. To conduct training on real-worl
d unpaired data current methods focus on synthesizing pseudo LR images to associ
ate unpaired images. However the realness and diversity of pseudo LR images are
vulnerable due to the large image space. In this paper we circumvent the difficu
lty of image generation and propose an alternative to build the connection betwe
en unpaired images in a compact proxy space. Specifically we first construct cou
pled HR and LR dictionaries and then encode HR and LR images into a common laten
t code space using these dictionaries. In addition we develop an autoencoder-bas
ed framework to couple these dictionaries during optimization by reconstructing
input HR and LR images. The coupled dictionaries enable our method to employ a s
hallow network architecture with only 18 layers to achieve efficient image SR. E
xtensive experiments show that our method (DictSR) can effectively model the LR-
to-HR mapping in coupled dictionaries and produces state-of-the-art performance
on benchmark datasets.
****
Template Free Reconstruction of Human-object Interaction with Procedural Interac
tion Generation

Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10003-10015
Reconstructing human-object interaction in 3D from a single RGB image is a challenging task and existing data driven methods do not generalize beyond the objects present in the carefully curated 3D interaction datasets. Capturing large-scale real data to learn strong interaction and 3D shape priors is very expensive due to the combinatorial nature of human-object interactions. In this paper we propose ProciGen (Procedural interaction Generation) a method to procedurally generate datasets with both plausible interaction and diverse object variation. We generate 1M+ human-object interaction pairs in 3D and leverage this large-scale data to train our HDM (Hierarchical Diffusion Model) a novel method to reconstruct interacting human and unseen object instances without any templates. Our HDM is an image-conditioned diffusion model that learns both realistic interaction and highly accurate human and object shapes. Experiments show that our HDM trained with ProciGen significantly outperforms prior methods that require template meshes and our dataset allows training methods with strong generalization ability to unseen object instances. Our code and data are released.
****
Deep Video Inverse Tone Mapping Based on Temporal Clues
Yuyao Ye, Ning Zhang, Yang Zhao, Hongbin Cao, Ronggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25995-26004
Inverse tone mapping (ITM) aims to reconstruct high dynamic range (HDR) radiance from low dynamic range (LDR) content. Although many deep image ITM methods can generate impressive results the field of video ITM is still to be explored. Processing video sequences by image ITM methods may cause temporal inconsistency. Besides they aren't able to exploit the potentially useful information in the temporal domain. In this paper we analyze the process of video filming and then propose a Global Sample and Local Propagate strategy to better find and utilize temporal clues. To better realize the proposed strategy we design a two-stage pipeline which includes modules named Incremental Clue Aggregation Module and Feature and Clue Propagation Module. They can align and fuse frames effectively under the condition of brightness changes and propagate features and temporal clues to all frames efficiently. Our temporal clues based video ITM method can recover realistic and temporal consistent results with high fidelity in over-exposed regions. Qualitative and quantitative experiments on public datasets show that the proposed method has significant advantages over existing methods.
****
NeRF-HuGS: Improved Neural Radiance Fields in Non-static Scenes Using Heuristics-Guided Segmentation
Jiahao Chen, Yipeng Qin, Lingjie Liu, Jiangbo Lu, Guanbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19436-19446
Neural Radiance Field (NeRF) has been widely recognized for its excellence in novel view synthesis and 3D scene reconstruction. However their effectiveness is inherently tied to the assumption of static scenes rendering them susceptible to undesirable artifacts when confronted with transient distractors such as moving objects or shadows. In this work we propose a novel paradigm namely "Heuristics-Guided Segmentation" (HuGS) which significantly enhances the separation of static scenes from transient distractors by harmoniously combining the strengths of hand-crafted heuristics and state-of-the-art segmentation models thus significantly transcending the limitations of previous solutions. Furthermore we delve into the meticulous design of heuristics introducing a seamless fusion of Structure-from-Motion (SfM)-based heuristics and color residual heuristics catering to a diverse range of texture profiles. Extensive experiments demonstrate the superiority and robustness of our method in mitigating transient distractors for NeRFs trained in non-static scenes. Project page: https://cnhaox.github.io/NeRF-HuGS/
****
Addressing Background Context Bias in Few-Shot Segmentation through Iterative Mo

dulation
Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3370-3379
Existing few-shot segmentation methods usually extract foreground prototypes from support images to guide query image segmentation. However different background contexts of support and query images can cause their foreground features to be misaligned. This phenomenon known as background context bias can hinder the effectiveness of support prototypes in guiding query image segmentation. In this work we propose a novel framework with an iterative structure to address this problem. In each iteration of the framework we first generate a query prediction based on a support foreground feature. Next we extract background context from the query image to modulate the support foreground feature thus eliminating the foreground feature misalignment caused by the different backgrounds. After that we design a confidence-biased attention to eliminate noise and cleanse information. By integrating these components through an iterative structure we create a novel network that can leverage the synergies between different modules to improve their performance in a mutually reinforcing manner. Through these carefully designed components and structures our network can effectively eliminate background context bias in few-shot segmentation thus achieving outstanding performance. We conduct extensive experiments on the PASCAL-5^ i and COCO-20^ i datasets and achieve state-of-the-art (SOTA) results which demonstrate the effectiveness of our approach.
****
Open-Vocabulary Video Anomaly Detection
Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18297-18307
Current video anomaly detection (VAD) approaches with weak supervisions are inherently limited to a closed-set setting and may struggle in open-world applications where there can be anomaly categories in the test data unseen during training. A few recent studies attempt to tackle a more realistic setting open-set VAD which aims to detect unseen anomalies given seen anomalies and normal videos. However such a setting focuses on predicting frame anomaly scores having no ability to recognize the specific categories of anomalies despite the fact that this ability is essential for building more informed video surveillance systems. This paper takes a step further and explores open-vocabulary video anomaly detection (OVVAD) in which we aim to leverage pre-trained large models to detect and categorize seen and unseen anomalies. To this end we propose a model that decouples OVVAD into two mutually complementary tasks - class-agnostic detection and class-specific classification - and jointly optimizes both tasks. Particularly we devise a semantic knowledge injection module to introduce semantic knowledge from large language models for the detection task and design a novel anomaly synthesis module to generate pseudo unseen anomaly videos with the help of large vision generation models for the classification task. These semantic knowledge and synthesis anomalies substantially extend our model's capability in detecting and categorizing a variety of seen and unseen anomalies. Extensive experiments on three widely-used benchmarks demonstrate our model achieves state-of-the-art performance on OVVAD task.
****
ODM: A Text-Image Further Alignment Pre-training Approach for Scene Text Detection and Spotting
Chen Duan, Pei Fu, Shan Guo, Qianyi Jiang, Xiaoming Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15587-15597
Abstract In recent years text-image joint pre-training techniques have shown promising results in various tasks. However in Optical Character Recognition (OCR) tasks aligning text instances with their corresponding text regions in images poses a challenge as it requires effective alignment between text and OCR-Text (referring to the text in images as OCR-Text to distinguish from the text in natura

l language) rather than a holistic understanding of the overall image content. I
n this paper we propose a new pre-training method called OCR-Text Destylization
Modeling (ODM) that transfers diverse styles of text found in images to a unifor
m style based on the text prompt. With ODM we achieve better alignment between t
ext and OCR-Text and enable pre-trained models to adapt to the complex and diver
se styles of scene text detection and spotting tasks. Additionally we have desig
ned a new labeling generation method specifically for ODM and combined it with o
ur proposed Text-Controller module to address the challenge of annotation costs
in OCR tasks allowing a larger amount of unlabeled data to participate in pre-tr
aining. Extensive experiments on multiple public datasets demonstrate that our m
ethod significantly improves performance and outperforms current pre-training me
thods in scene text detection and spotting tasks. Code is available at https://g
ithub.com/PriNing/ODM.
****

TiNO-Edit: Timestep and Noise Optimization for Robust Diffusion-Based Image Edit
ing
Sherry X Chen, Yaron Vaxman, Elad Ben Baruch, David Asulin, Aviad Moreshet, Kuo-
Chin Lien, Misha Sra, Pradeep Sen; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2024, pp. 6337-6346
Despite many attempts to leverage pre-trained text-to-image models (T2I) like St
able Diffusion (SD) for controllable image editing producing good predictable re
sults remains a challenge. Previous approaches have focused on either fine-tunin
g pre-trained T2I models on specific datasets to generate certain kinds of image
s (e.g. with a specific object or person) or on optimizing the weights text prom
pts and/or learning features for each input image in an attempt to coax the imag
e generator to produce the desired result. However these approaches all have sho
rtcomings and fail to produce good results in a predictable and controllable man
ner. To address this problem we present TiNO-Edit an SD-based method that focuse
s on optimizing the noise patterns and diffusion timesteps during editing someth
ing previously unexplored in the literature. With this simple change we are able
 to generate results that both better align with the original images and reflect
 the desired result. Furthermore we propose a set of new loss functions that ope
rate in the latent domain of SD greatly speeding up the optimization when compar
ed to prior losses which operate in the pixel domain. Our method can be easily a
pplied to variations of SD including Textual Inversion and DreamBooth that encod
e new concepts and incorporate them into the edited results. We present a host o
f image-editing capabilities enabled by our approach. Our code is publicly avail
able at https://github.com/SherryXTChen/TiNO-Edit.
****

Epistemic Uncertainty Quantification For Pre-Trained Neural Networks
Hanjing Wang, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2024, pp. 11052-11061
Epistemic uncertainty quantification (UQ) identifies where models lack knowledge
. Traditional UQ methods often based on Bayesian neural networks are not suitabl
e for pre-trained non-Bayesian models. Our study addresses quantifying epistemic
 uncertainty for any pre-trained model which does not need the original training
 data or model modifications and can ensure broad applicability regardless of ne
twork architectures or training techniques. Specifically we propose a gradient-b
ased approach to assess epistemic uncertainty analyzing the gradients of outputs
 relative to model parameters and thereby indicating necessary model adjustments
 to accurately represent the inputs. We first explore theoretical guarantees of
gradient-based methods for epistemic UQ questioning the view that this uncertain
ty is only calculable through differences between multiple models. We further im
prove gradient-driven UQ by using class-specific weights for integrating gradien
ts and emphasizing distinct contributions from neural network layers. Additional
ly we enhance UQ accuracy by combining gradient and perturbation methods to refi
ne the gradients. We evaluate our approach on out-of-distribution detection unce
rtainty calibration and active learning demonstrating its superiority over curre
nt state-of-the-art UQ methods for pre-trained models.
****

Diffusion-ES: Gradient-free Planning with Diffusion for Autonomous and Instruction-guided Driving
Brian Yang, Huangyuan Su, Nikolaos Gkanatsios, Tsung-Wei Ke, Ayush Jain, Jeff Schneider, Katerina Fragkiadaki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15342-15353
Diffusion models excel at modeling complex and multimodal trajectory distributions for decision-making and control. Reward-gradient guided denoising has been recently proposed to generate trajectories that maximize both a differentiable reward function and the likelihood under the data distribution captured by a diffusion model. Reward-gradient guided denoising requires a differentiable reward function fitted to both clean and noised samples limiting its applicability as a general trajectory optimizer. In this paper we propose DiffusionES a method that combines gradient-free optimization with trajectory denoising to optimize black-box non-differentiable objectives while staying in the data manifold. Diffusion-ES samples trajectories during evolutionary search from a diffusion model and scores them using a black-box reward function. It mutates high-scoring trajectories using a truncated diffusion process that applies a small number of noising and denoising steps allowing for much more efficient exploration of the solution space. We show that DiffusionES achieves state-of-the-art performance on nuPlan an established closed-loop planning benchmark for autonomous driving. Diffusion-ES outperforms existing sampling-based planners reactive deterministic or diffusion-based policies and reward-gradient guidance. Additionally we show that unlike prior guidance methods our method can optimize non-differentiable language-shaped reward functions generated by few-shot LLM prompting. When guided by a human teacher that issues instructions to follow our method can generate novel highly complex behaviors such as aggressive lane weaving which are not present in the training data. This allows us to solve the hardest nuPlan scenarios which are beyond the capabilities of existing trajectory optimization methods and driving policies.
****
AdaShift: Learning Discriminative Self-Gated Neural Feature Activation With an Adaptive Shift Factor
Sudong Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5947-5956
Nonlinearities are decisive in neural representation learning. Traditional Activation (Act) functions impose fixed inductive biases on neural networks with oriented biological intuitions. Recent methods leverage self-gated curves to compensate for the rigid traditional Act paradigms in fitting flexibility. However substantial improvements are still impeded by the norm-induced mismatched feature re-calibrations (see Section 1) i.e. the actual importance of a feature can be inconsistent with its explicit intensity such that violates the basic intention of a direct self-gated feature re-weighting. To address this problem we propose to learn discriminative neural feature Act with a novel prototype namely AdaShift which enhances typical self-gated Act by incorporating an adaptive shift factor into the re-weighting function of Act. AdaShift casts dynamic translations on the inputs of a re-weighting function by exploiting comprehensive feature-filter context cues of different ranges in a simple yet effective manner. We obtain the new intuitions of AdaShift by rethinking the feature-filter relationships from a common Softmax-based classification and by generalizing the new observations to a common learning layer that encodes features with updatable filters. Our practical AdaShifts built upon the new Act prototype demonstrate significant improvements to the popular/SOTA Act functions on different vision benchmarks. By simply replacing ReLU with AdaShifts ResNets can match advanced Transformer counterparts (e.g. ResNet-50 vs. Swin-T) with lower cost and fewer parameters.
****
SCEdit: Efficient and Controllable Image Diffusion Generation via Skip Connection Editing
Zeyinzi Jiang, Chaojie Mao, Yulin Pan, Zhen Han, Jingfeng Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8995-9004

Image diffusion models have been utilized in various tasks such as text-to-image generation and controllable image synthesis. Recent research has introduced tuning methods that make subtle adjustments to the original models yielding promising results in specific adaptations of foundational generative diffusion models. Rather than modifying the main backbone of the diffusion model we delve into the role of skip connection in U-Net and reveal that hierarchical features aggregating long-distance information across encoder and decoder make a significant impact on the content and quality of image generation. Based on the observation we propose an efficient generative tuning framework dubbed SCEdit which integrates and edits Skip Connection using a lightweight tuning module named SC-Tuner. Furthermore the proposed framework allows for straightforward extension to controllable image synthesis by injecting different conditions with Controllable SC-Tuner simplifying and unifying the network design for multi-condition inputs. Our SCEdit substantially reduces training parameters memory usage and computational expense due to its lightweight tuners with backward propagation only passing to the decoder blocks. Extensive experiments conducted on text-to-image generation and controllable image synthesis tasks demonstrate the superiority of our method in terms of efficiency and performance. Project page: https://scedit.github.io/.
****
MRC-Net: 6-DoF Pose Estimation with MultiScale Residual Correlation
Yuelong Li, Yafei Mao, Raja Bala, Sunil Hadap; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10476-10486
We propose a single-shot approach to determining 6-DoF pose of an object with available 3D computer-aided design (CAD) model from a single RGB image. Our method dubbed MRC-Net comprises two stages. The first performs pose classification and renders the 3D object in the classified pose. The second stage performs regression to predict fine-grained residual pose within class. Connecting the two stages is a novel multi-scale residual correlation (MRC) layer that captures high-and-low level correspondences between the input image and rendering from first stage. MRC-Net employs a Siamese network with shared weights between both stages to learn embeddings for input and rendered images. To mitigate ambiguity when predicting discrete pose class labels on symmetric objects we use soft probabilistic labels to define pose class in the first stage. We demonstrate state-of-the-art accuracy outperforming all competing RGB-based methods on four challenging BOP benchmark datasets: T-LESS LM-O YCB-V and ITODD. Our method is non-iterative and requires no complex post-processing. Our code and pretrained models are available at https://github.com/amzn/mrc-net-6d-pose
****
MonoCD: Monocular 3D Object Detection with Complementary Depths
Longfei Yan, Pei Yan, Shengzhou Xiong, Xuanyu Xiang, Yihua Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10248-10257
Monocular 3D object detection has attracted widespread attention due to its potential to accurately obtain object 3D localization from a single image at a low cost. Depth estimation is an essential but challenging subtask of monocular 3D object detection due to the ill-posedness of 2D to 3D mapping. Many methods explore multiple local depth clues such as object heights and keypoints and then formulate the object depth estimation as an ensemble of multiple depth predictions to mitigate the insufficiency of single-depth information. However the errors of existing multiple depths tend to have the same sign which hinders them from neutralizing each other and limits the overall accuracy of combined depth. To alleviate this problem we propose to increase the complementarity of depths with two novel designs. First we add a new depth prediction branch named complementary depth that utilizes global and efficient depth clues from the entire image rather than the local clues to reduce the correlation of depth predictions. Second we propose to fully exploit the geometric relations between multiple depth clues to achieve complementarity in form. Benefiting from these designs our method achieves higher complementarity. Experiments on the KITTI benchmark demonstrate that our method achieves state-of-the-art performance without introducing extra data. In addition complementary depth can also be a lightweight and plug-and-play module

to boost multiple existing monocular 3d object detectors. Code is available at https://github.com/elvintanhust/MonoCD.
****

ImageNet-D: Benchmarking Neural Network Robustness on Diffusion Synthetic Object
Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, Chengzhi Mao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21752-21762
We establish rigorous benchmarks for visual perception robustness. Synthetic images such as ImageNet-C ImageNet-9 and Stylized ImageNet provide specific type of evaluation over synthetic corruptions backgrounds and textures yet those robustness benchmarks are restricted in specified variations and have low synthetic quality. In this work we introduce generative model as a data source for synthesizing hard images that benchmark deep models' robustness. Leveraging diffusion models we are able to generate images with more diversified backgrounds textures and materials than any prior work where we term this benchmark as ImageNet-D. Experimental results show that ImageNet-D results in a significant accuracy drop to a range of vision models from the standard ResNet visual classifier to the latest foundation models like CLIP and MiniGPT-4 significantly reducing their accuracy by up to 60%. Our work suggests that diffusion models can be an effective source to test vision models. The code and dataset are available at https://github.com/chenshuang-zhang/imagenet_d.
****

Consistent3D: Towards Consistent High-Fidelity Text-to-3D Generation with Deterministic Sampling Prior
Zike Wu, Pan Zhou, Xuanyu Yi, Xiaoding Yuan, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9892-9902
Score distillation sampling (SDS) and its variants have greatly boosted the development of text-to-3D generation but are vulnerable to geometry collapse and poor textures yet. To solve this issue we first deeply analyze the SDS and find that its distillation sampling process indeed corresponds to the trajectory sampling of a stochastic differential equation (SDE): SDS samples along an SDE trajectory to yield a less noisy sample which then serves as a guidance to optimize a 3D model. However the randomness in SDE sampling often leads to a diverse and unpredictable sample which is not always less noisy and thus is not a consistently correct guidance explaining the vulnerability of SDS. Since for any SDE there always exists an ordinary differential equation (ODE) whose trajectory sampling can deterministically and consistently converge to the desired target point as the SDE we propose a novel and effective "Consistent3D" method that explores the ODE deterministic sampling prior for text-to-3D generation. Specifically at each training iteration given a rendered image by a 3D model we first estimate its desired 3D score function by a pre-trained 2D diffusion model and build an ODE for trajectory sampling. Next we design a consistency distillation sampling loss which samples along the ODE trajectory to generate two adjacent samples and uses the less noisy sample to guide another more noisy one for distilling the deterministic prior into the 3D model. Experimental results show the efficacy of our Consistent3D in generating high-fidelity and diverse 3D objects and large-scale scenes as shown in Fig. 1. The codes are available at https://github.com/sail-sg/Consistent3D.
****

ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation
Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, Hao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18061-18070
Robot manipulation relies on accurately predicting contact points and end-effector directions to ensure successful operation. However learning-based robot manipulation trained on a limited category within a simulator often struggles to achieve generalizability especially when confronted with extensive categories. Therefore we introduce an innovative approach for robot manipulation that leverages t

he robust reasoning capabilities of Multimodal Large Language Models (MLLMs) to enhance the stability and generalization of manipulation. By fine-tuning the injected adapters we preserve the inherent common sense and reasoning ability of the MLLMs while equipping them with the ability for manipulation. The fundamental insight lies in the introduced fine-tuning paradigm encompassing object category understanding affordance prior reasoning and object-centric pose prediction to stimulate the reasoning ability of MLLM in manipulation. During inference our approach utilizes an RGB image and text prompt to predict the end effector's pose in chain of thoughts. After the initial contact is established an active impedance adaptation policy is introduced to plan the upcoming waypoints in a closed-loop manner. Moreover in real world we design a test-time adaptation (TTA) strategy for manipulation to enable the model better adapt to the current real-world scene configuration. Experiments in simulator and real-world show the promising performance of ManipLLM. More details and demonstrations can be found at https://sites.google.com/view/manipllm.
****
BA-SAM: Scalable Bias-Mode Attention Mask for Segment Anything Model
Yiran Song, Qianyu Zhou, Xiangtai Li, Deng-Ping Fan, Xuequan Lu, Lizhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3162-3173
In this paper we address the challenge of image resolution variation for the Segment Anything Model (SAM). SAM known for its zero-shot generalizability exhibits a performance degradation when faced with datasets with varying image sizes. Previous approaches tend to resize the image to a fixed size or adopt structure modifications hindering the preservation of SAM's rich prior knowledge. Besides such task-specific tuning necessitates a complete retraining of the model which is cost-expensive and unacceptable for deployment in the downstream tasks. In this paper we reformulate this challenge as a length extrapolation problem where token sequence length varies while maintaining a consistent patch size for images with different sizes. To this end we propose a Scalable Bias-Mode Attention Mask (BA-SAM) to enhance SAM's adaptability to varying image resolutions while eliminating the need for structure modifications. Firstly we introduce a new scaling factor to ensure consistent magnitude in the attention layer's dot product values when the token sequence length changes. Secondly we present a bias-mode attention mask that allows each token to prioritize neighboring information mitigating the impact of untrained distant information. Our BA-SAM demonstrates efficacy in two scenarios: zero-shot and fine-tuning. Extensive evaluation of diverse datasets including DIS5K DUTS ISIC COD10K and COCO reveals its ability to significantly mitigate performance degradation in the zero-shot setting and achieve state-of-the-art performance with minimal fine-tuning. Furthermore we propose a generalized model and benchmark showcasing BA-SAM's generalizability across all four datasets simultaneously.
****
Text-Enhanced Data-free Approach for Federated Class-Incremental Learning
Minh-Tuan Tran, Trung Le, Xuan-May Le, Mehrtash Harandi, Dinh Phung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23870-23880
Federated Class-Incremental Learning (FCIL) is an underexplored yet pivotal issue involving the dynamic addition of new classes in the context of federated learning. In this field Data-Free Knowledge Transfer (DFKT) plays a crucial role in addressing catastrophic forgetting and data privacy problems. However prior approaches lack the crucial synergy between DFKT and the model training phases causing DFKT to encounter difficulties in generating high-quality data from a non-anchored latent space of the old task model. In this paper we introduce LANDER (Label Text Centered Data-Free Knowledge Transfer) to address this issue by utilizing label text embeddings (LTE) produced by pretrained language models. Specifically during the model training phase our approach treats LTE as anchor points and constrains the feature embeddings of corresponding training samples around them enriching the surrounding area with more meaningful information. In the DFKT phase by using these LTE anchors LANDER can synthesize more meaningful samples ther

eby effectively addressing the forgetting problem. Additionally instead of tight ly constraining embeddings toward the anchor the Bounding Loss is introduced to encourage sample embeddings to remain flexible within a defined radius. This app roach preserves the natural differences in sample embeddings and mitigates the e mbedding overlap caused by heterogeneous federated settings. Extensive experimen ts conducted on CIFAR100 Tiny-ImageNet and ImageNet demonstrate that LANDER sign ificantly outperforms previous methods and achieves state-of-the-art performance in FCIL. The code is available at https://github.com/tmtuan1307/lander.
****
Deciphering 'What' and 'Where' Visual Pathways from Spectral Clustering of Layer -Distributed Neural Representations
Xiao Zhang, David Yunis, Michael Maire; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4165-4175
We present an approach for analyzing grouping information contained within a neu ral network's activations permitting extraction of spatial layout and semantic s egmentation from the behavior of large pre-trained vision models. Unlike prior w ork our method conducts a wholistic analysis of a network's activation state lev eraging features from all layers and obviating the need to guess which part of t he model contains relevant information. Motivated by classic spectral clustering we formulate this analysis in terms of an optimization objective involving a se t of affinity matrices each formed by comparing features within a different laye r. Solving this optimization problem using gradient descent allows our technique to scale from single images to dataset-level analysis including in the latter b oth intra- and inter-image relationships. Analyzing a pre-trained generative tra nsformer provides insight into the computational strategy learned by such models . Equating affinity with key-query similarity across attention layers yields eig envectors encoding scene spatial layout whereas defining affinity by value vecto r similarity yields eigenvectors encoding object identity. This result suggests that key and query vectors coordinate attentional information flow according to spatial proximity (a `where' pathway) while value vectors refine a semantic cate gory representation (a `what' pathway).
****
GLaMM: Pixel Grounding Large Multimodal Model
Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hi sham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, Fahad S. Khan; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13009-13018
Large Multimodal Models (LMMs) extend Large Language Models to the vision domain . Initial LMMs used holistic images and text prompts to generate ungrounded text ual responses. Recently region-level LMMs have been used to generate visually gr ounded responses. However they are limited to only referring to a single object category at a time require users to specify the regions or cannot offer dense pi xel-wise object grounding. In this work we present Grounding LMM (GLaMM) the fir st model that can generate natural language responses seamlessly intertwined wit h corresponding object segmentation masks. GLaMM not only grounds objects appear ing in the conversations but is flexible enough to accept both textual and optio nal visual prompts (region of interest) as input. This empowers users to interac t with the model at various levels of granularity both in textual and visual dom ains. Due to the lack of standard benchmarks for the novel setting of visually G rounded Conversation Generation (GCG) we introduce a comprehensive evaluation pr otocol with our curated grounded conversations. Our proposed GCG task requires d ensely grounded concepts in natural scenes at a large-scale. To this end we prop ose a densely annotated Grounding-anything Dataset (GranD) using our proposed au tomated annotation pipeline that encompasses 7.5M unique concepts grounded in a total of 810M regions available with segmentation masks. Besides GCG GLaMM also performs effectively on several downstream tasks e.g. referring expression segme ntation image and region-level captioning and vision-language conversations.
****
Incremental Residual Concept Bottleneck Models
Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, Yuwang Wang;

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11030-11040
Concept Bottleneck Models (CBMs) map the black-box visual representations extracted by deep neural networks onto a set of interpretable concepts and use the concepts to make predictions enhancing the transparency of the decision-making process. Multimodal pre-trained models can match visual representations with textual concept embeddings allowing for obtaining the interpretable concept bottleneck without the expertise concept annotations. Recent research has focused on the concept bank establishment and the high-quality concept selection. However it is challenging to construct a comprehensive concept bank through humans or large language models which severely limits the performance of CBMs. In this work we propose the Incremental Residual Concept Bottleneck Model (Res-CBM) to address the challenge of concept completeness. Specifically the residual concept bottleneck model employs a set of optimizable vectors to complete missing concepts then the incremental concept discovery module converts the complemented vectors with unclear meanings into potential concepts in the candidate concept bank. Our approach can be applied to any user-defined concept bank as a post-hoc processing method to enhance the performance of any CBMs. Furthermore to measure the descriptive efficiency of CBMs the Concept Utilization Efficiency (CUE) metric is proposed. Experiments show that the Res-CBM outperforms the current state-of-the-art methods in terms of both accuracy and efficiency and achieves comparable performance to black-box models across multiple datasets.
****
SPOC: Imitating Shortest Paths in Simulation Enables Effective Navigation and Manipulation in the Real World
Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, Ranjay Krishna, Dustin Schwenk, Eli VanderBilt, Aniruddha Kembhavi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16238-16250
Reinforcement learning (RL) with dense rewards and imitation learning (IL) with human-generated trajectories are the most widely used approaches for training modern embodied agents. RL requires extensive reward shaping and auxiliary losses and is often too slow and ineffective for long-horizon tasks. While IL with human supervision is effective collecting human trajectories at scale is extremely expensive. In this work we show that imitating shortest-path planners in simulation produces agents that given a language instruction can proficiently navigate explore and manipulate objects in both simulation and in the real world using only RGB sensors (no depth map or GPS coordinates). This surprising result is enabled by our end-to-end transformer-based SPOC architecture powerful visual encoders paired with extensive image augmentation and the dramatic scale and diversity of our training data: millions of frames of shortest-path-expert trajectories collected inside approximately 200000 procedurally generated houses containing 40000 unique 3D assets. Our models data training code and newly proposed 10-task benchmarking suite CHORES are available at https://spoc-robot.github.io/.
****
Real-Time Exposure Correction via Collaborative Transformations and Adaptive Sampling
Ziwen Li, Feng Zhang, Meng Cao, Jinpu Zhang, Yuanjie Shao, Yuehuan Wang, Nong Sang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2984-2994
Most of the previous exposure correction methods learn dense pixel-wise transformations to achieve promising results but consume huge computational resources. Recently Learnable 3D lookup tables (3D LUTs) have demonstrated impressive performance and efficiency for image enhancement. However these methods can only perform global transformations and fail to finely manipulate local regions. Moreover they uniformly downsample the input image which loses the rich color information and limits the learning of color transformation capabilities. In this paper we present a collaborative transformation framework (CoTF) for real-time exposure correction which integrates global transformation with pixel-wise transformations

in an efficient manner. Specifically the global transformation adjusts the overall appearance using image-adaptive 3D LUTs to provide decent global contrast and sharp details while the pixel transformation compensates for local context. Then a relation-aware modulation module is designed to combine these two components effectively. In addition we propose an adaptive sampling strategy to preserve more color information by predicting the sampling intervals thus providing higher quality input data for the learning of 3D LUTs. Extensive experiments demonstrate that our method can process high-resolution images in real-time on GPUs while achieving comparable performance against current state-of-the-art methods. The code is available at https://github.com/HUST-IAL/CoTF.
****
Lodge: A Coarse to Fine Diffusion Network for Long Dance Generation Guided by the Characteristic Dance Primitives
Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, Xiu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1524-1534
We propose Lodge a network capable of generating extremely long dance sequences conditioned on given music. We design Lodge as a two-stage coarse to fine diffusion architecture and propose the characteristic dance primitives that possess significant expressiveness as intermediate representations between two diffusion models. The first stage is global diffusion which focuses on comprehending the coarse-level music-dance correlation and production characteristic dance primitives. In contrast the second-stage is the local diffusion which parallelly generates detailed motion sequences under the guidance of the dance primitives and choreographic rules. In addition we propose a Foot Refine Block to optimize the contact between the feet and the ground enhancing the physical realism of the motion. Code available at https://li-ronghui.github.io/lodge
****
UDiFF: Generating Conditional Unsigned Distance Fields with Optimal Wavelet Diffusion
Junsheng Zhou, Weiqi Zhang, Baorui Ma, Kanle Shi, Yu-Shen Liu, Zhizhong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21496-21506
Diffusion models have shown remarkable results for image generation editing and inpainting. Recent works explore diffusion models for 3D shape generation with neural implicit functions i.e. signed distance function and occupancy function. However they are limited to shapes with closed surfaces which prevents them from generating diverse 3D real-world contents containing open surfaces. In this work we present UDiFF a 3D diffusion model for unsigned distance fields (UDFs) which is capable to generate textured 3D shapes with open surfaces from text conditions or unconditionally. Our key idea is to generate UDFs in spatial-frequency domain with an optimal wavelet transformation which produces a compact representation space for UDF generation. Specifically instead of selecting an appropriate wavelet transformation which requires expensive manual efforts and still leads to large information loss we propose a data-driven approach to learn the optimal wavelet transformation for UDFs. We evaluate UDiFF to show our advantages by numerical and visual comparisons with the latest methods on widely used benchmarks. Page: https://weiqi-zhang.github.io/UDiFF.
****
LoCoNet: Long-Short Context Network for Active Speaker Detection
Xizi Wang, Feng Cheng, Gedas Bertasius; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18462-18472
Active Speaker Detection (ASD) aims to identify who is speaking in each frame of a video. Solving ASD involves using audio and visual information in two complementary contexts: long-term intra-speaker context models the temporal dependencies of the same speaker while short-term inter-speaker context models the interactions of speakers in the same scene. Motivated by these observations we propose LoCoNet a simple but effective Long-Short Context Network that leverages Long-term Intra-speaker Modeling (LIM) and Short-term Inter-speaker Modeling (SIM) in an interleaved manner. LIM employs self-attention for long-range temporal dependen

cies modeling and cross-attention for audio-visual interactions modeling. SIM in corporates convolutional blocks that capture local patterns for short-term inter -speaker context. Experiments show that LoCoNet achieves state-of-the-art perfor mance on multiple datasets with 95.2% (+0.3%) mAP on AVA-ActiveSpeaker 97.2% (+2 .7%) mAP on Talkies and 68.4% (+7.7%) mAP on Ego4D. Moreover in challenging case s where multiple speakers are present LoCoNet outperforms previous state-of-the- art methods by 3.0% mAP on AVA-ActiveSpeaker. The code is available at https://g ithub.com/SJTUwxz/LoCoNet_ASD.
****
D3still: Decoupled Differential Distillation for Asymmetric Image Retrieval
Yi Xie, Yihong Lin, Wenjie Cai, Xuemiao Xu, Huaidong Zhang, Yong Du, Shengfeng H e; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2024, pp. 17181-17190
Existing methods for asymmetric image retrieval employ a rigid pairwise similari ty constraint between the query network and the larger gallery network. However these one-to-one constraint approaches often fail to maintain retrieval order co nsistency especially when the query network has limited representational capacit y. To overcome this problem we introduce the Decoupled Differential Distillation (D3still) framework. This framework shifts from absolute one-to-one supervision to optimizing the relational differences in pairwise similarities produced by t he query and gallery networks thereby preserving a consistent retrieval order ac ross both networks. Our method involves computing a pairwise similarity differen tial matrix within the gallery domain which is then decomposed into three compon ents: feature representation knowledge inconsistent pairwise similarity differen tial knowledge and consistent pairwise similarity differential knowledge. This s trategic decomposition aligns the retrieval ranking of the query network with th e gallery network effectively. Extensive experiments on various benchmark datase ts reveal that D3still surpasses state-of-the-art methods in asymmetric image re trieval. Code is available at https://github.com/SCY-X/D3still.
****
Transcending Forgery Specificity with Latent Space Augmentation for Generalizabl e Deepfake Detection
Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, Baoyuan Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8984-8994
Deepfake detection faces a critical generalization hurdle with performance deter iorating when there is a mismatch between the distributions of training and test ing data. A broadly received explanation is the tendency of these detectors to b e overfitted to forgery-specific artifacts rather than learning features that ar e widely applicable across various forgeries. To address this issue we propose a simple yet effective detector called LSDA (\underline L atent \underline S pace \underline D ata \underline A ugmentation) which is based on a heuristic idea: representations with a wider variety of forgeries should be able to learn a more generalizable decision boundary thereby mitigating the overfitting of method-sp ecific features (see Fig. 1). Following this idea we propose to enlarge the forg ery space by constructing and simulating variations within and across forgery fe atures in the latent space. This approach encompasses the acquisition of enriche d domain-specific features and the facilitation of smoother transitions between different forgery types effectively bridging domain gaps. Our approach culminate s in refining a binary classifier that leverages the distilled knowledge from th e enhanced features striving for a generalizable deepfake detector. Comprehensiv e experiments show that our proposed method is surprisingly effective and transc ends state-of-the-art detectors across several widely used benchmarks.
****
Scaling Laws of Synthetic Images for Model Training ... for Now
Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, Yonglong Ti an; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2024, pp. 7382-7392
Recent significant advances in text-to-image models unlock the possibility of tr aining vision systems using synthetic images potentially overcoming the difficul

ty of collecting curated data at scale. It is unclear however how these models behave at scale as more synthetic data is added to the training set. In this paper we study the scaling laws of synthetic images generated by state of the art text-to-image models for the training of supervised models: image classifiers with label supervision and CLIP with language supervision. We identify several factors including text prompts classifier-free guidance scale and types of text-to-image models that significantly affect scaling behavior. After tuning these factors we observe that synthetic images demonstrate a scaling trend similar to but slightly less effective than real images in CLIP training while they significantly underperform in scaling when training supervised image classifiers. Our analysis indicates that the main reason for this underperformance is the inability of off-the-shelf text-to-image models to generate certain concepts a limitation that significantly impairs the training of image classifiers. Our findings also suggest that scaling synthetic data can be particularly effective in scenarios such as: (1) when there is a limited supply of real images for a supervised problem (e.g. fewer than 0.5 million images in ImageNet) (2) when the evaluation dataset diverges significantly from the training data indicating the out-of-distribution scenario or (3) when synthetic data is used in conjunction with real images as demonstrated in the training of CLIP models.
****

Towards Large-scale 3D Representation Learning with Multi-dataset Point Prompt Training

Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, Hengshuang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19551-19562

The rapid advancement of deep learning models is often attributed to their ability to leverage massive training data. In contrast such privilege has not yet fully benefited 3D deep learning mainly due to the limited availability of large-scale 3D datasets. Merging multiple available data sources and letting them collaboratively train a single model is a potential solution. However due to the large domain gap between 3D point cloud datasets such mixed supervision could adversely affect the model's performance and lead to degenerated performance (i.e. negative transfer) compared to single-dataset training. In view of this challenge we introduce Point Prompt Training (PPT) a novel framework for multi-dataset synergistic learning in the context of 3D representation learning that supports multiple pre-training paradigms. Based on this framework we propose Prompt-driven Normalization which adapts the model to different datasets with domain-specific prompts and Language-guided Categorical Alignment that decently unifies the multiple-dataset label spaces by leveraging the relationship between label text. Extensive experiments verify that PPT can overcome the negative transfer associated with synergistic learning and produce generalizable representations. Notably it achieves state-of-the-art performance on each dataset using a single weight-shared model with supervised multi-dataset training. Moreover when served as a pre-training framework it outperforms other pre-training approaches regarding representation quality and attains remarkable state-of-the-art performance across over ten diverse downstream tasks spanning both indoor and outdoor 3D scenarios.
****

Learning Triangular Distribution in Visual World

Ping Chen, Xingpeng Zhang, Chengtao Zhou, Dichao Fan, Peng Tu, Le Zhang, Yanlin Qian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11019-11029

Convolution neural network is successful in pervasive vision tasks including label distribution learning which usually takes the form of learning an injection from the non-linear visual features to the well-defined labels. However how the discrepancy between features is mapped to the label discrepancy is ambient and its correctness is not guaranteed.To address these problems we study the mathematical connection between feature and its label presenting a general and simple framework for label distribution learning. We propose a so-called Triangular Distribution Transform (TDT) to build an injective function between feature and label guaranteeing that any symmetric feature discrepancy linearly reflects the differ

ence between labels. The proposed TDT can be used as a plug-in in mainstream backbone networks to address different label distribution learning tasks. Experiments on Facial Age Recognition Illumination Chromaticity Estimation and Aesthetics assessment show that TDT achieves on-par or better results than the prior arts.
****
State Space Models for Event Cameras
Nikola Zubic, Mathias Gehrig, Davide Scaramuzza; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5819-5828
Today state-of-the-art deep neural networks that process event-camera data first convert a temporal window of events into dense grid-like input representations. As such they exhibit poor generalizability when deployed at higher inference frequencies (i.e. smaller temporal windows) than the ones they were trained on. We address this challenge by introducing state-space models (SSMs) with learnable timescale parameters to event-based vision. This design adapts to varying frequencies without the need to retrain the network at different frequencies. Additionally we investigate two strategies to counteract aliasing effects when deploying the model at higher frequencies. We comprehensively evaluate our approach against existing methods based on RNN and Transformer architectures across various benchmarks including Gen1 and 1 Mpx event camera datasets. Our results demonstrate that SSM-based models train 33% faster and also exhibit minimal performance degradation when tested at higher frequencies than the training input. Traditional RNN and Transformer models exhibit performance drops of more than 20 mAP with SSMs having a drop of 3.31 mAP highlighting the effectiveness of SSMs in event-based vision tasks.
****
EmbodiedScan: A Holistic Multi-Modal 3D Perception Suite Towards Embodied AI
Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, Jiangmiao Pang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19757-19767
In the realm of computer vision and robotics embodied agents are expected to explore their environment and carry out human instructions. This necessitates the ability to fully understand 3D scenes given their first-person observations and contextualize them into language for interaction. However traditional research focuses more on scene-level input and output setups from a global view. To address the gap we introduce EmbodiedScan a multi-modal ego-centric 3D perception dataset and benchmark for holistic 3D scene understanding. It encompasses over 5k scans encapsulating 1M ego-centric RGB-D views 1M language prompts 160k 3D-oriented boxes spanning over 760 categories some of which partially align with LVIS and dense semantic occupancy with 80 common categories. Building upon this database we introduce a baseline framework named Embodied Perceptron. It is capable of processing an arbitrary number of multi-modal inputs and demonstrates remarkable 3D perception capabilities both within the two series of benchmarks we set up i.e. fundamental 3D perception tasks and language-grounded tasks and in the wild.
****
SHINOBI: Shape and Illumination using Neural Object Decomposition via BRDF Optimization In-the-wild
Andreas Engelhardt, Amit Raj, Mark Boss, Yunzhi Zhang, Abhishek Kar, Yuanzhen Li, Deqing Sun, Ricardo Martin Brualla, Jonathan T. Barron, Hendrik P. A. Lensch, Varun Jampani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19636-19646
We present SHINOBI an end-to-end framework for the reconstruction of shape material and illumination from object images captured with varying lighting pose and background. Inverse rendering of an object based on unconstrained image collections is a long-standing challenge in computer vision and graphics and requires a joint optimization over shape radiance and pose. We show that an implicit shape representation based on a multi-resolution hash encoding enables faster and robust shape reconstruction with joint camera alignment optimization that outperforms prior work. Further to enable the editing of illumination and object reflectance (i.e. material) we jointly optimize BRDF and illumination together with the o

bject's shape. Our method is class-agnostic and works on in-the-wild image colle ctions of objects to produce relightable 3D assets for several use cases such as AR/VR movies games etc.
****
ES3: Evolving Self-Supervised Learning of Robust Audio-Visual Speech Representat ions
Yuanhang Zhang, Shuang Yang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2706 9-27079
We propose a novel strategy ES3 for self-supervised learning of robust audio-vis ual speech representations from unlabeled talking face videos. While many recent approaches for this task primarily rely on guiding the learning process using t he audio modality alone to capture information shared between audio and video we reframe the problem as the acquisition of shared unique (modality-specific) and synergistic speech information to address the inherent asymmetry between the mo dalities. Based on this formulation we propose a novel "evolving" strategy that progressively builds joint audio-visual speech representations that are strong f or both uni-modal (audio & visual) and bi-modal (audio-visual) speech. First we leverage the more easily learnable audio modality to initialize audio and visual representations by capturing audio-unique and shared speech information. Next w e incorporate video-unique speech information and bootstrap the audio-visual rep resentations on top of the previously acquired shared knowledge. Finally we maxi mize the total audio-visual speech information including synergistic information to obtain robust and comprehensive representations. We implement ES3 as a simpl e Siamese framework and experiments on both English benchmarks and a newly contr ibuted large-scale Mandarin dataset show its effectiveness. In particular on LRS 2-BBC our smallest model is on par with SoTA models with only 1/2 parameters and 1/8 unlabeled data (223h).
****
TeTriRF: Temporal Tri-Plane Radiance Fields for Efficient Free-Viewpoint Video
Minye Wu, Zehao Wang, Georgios Kouros, Tinne Tuytelaars; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 648 7-6496
Neural Radiance Fields (NeRF) revolutionize the realm of visual media by providi ng photorealistic Free-Viewpoint Video (FVV) experiences offering viewers unpara lleled immersion and interactivity. However the technology's significant storage requirements and the computational complexity involved in generation and render ing currently limit its broader application. To close this gap this paper presen ts Temporal Tri-Plane Radiance Fields (TeTriRF) a novel technology that signific antly reduces the storage size for Free-Viewpoint Video (FVV) while maintaining low-cost generation and rendering. TeTriRF introduces a hybrid representation wi th tri-planes and voxel grids to support scaling up to long-duration sequences a nd scenes with complex motions or rapid changes. We propose a group training sch eme tailored to achieving high training efficiency and yielding temporally consi stent low-entropy scene representations on feature domain. Leveraging these prop erties of the representations we introduce a compression pipeline with off-the-s helf video codecs achieving an order of magnitude less storage size compared to the state-of-the-art. Our experiments demonstrate that TeTriRF can achieve compe titive quality with a higher compression rate.
****
Motion2VecSets: 4D Latent Vector Set Diffusion for Non-rigid Shape Reconstructio n and Tracking
Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, Jiapeng Tang; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20496-20506
We introduce Motion2VecSets a 4D diffusion model for dynamic surface reconstruct ion from point cloud sequences. While existing state-of-the-art methods have dem onstrated success in reconstructing non-rigid objects using neural field represe ntations conventional feed-forward networks encounter challenges with ambiguous observations from noisy partial or sparse point clouds. To address these challen

ges we introduce a diffusion model that explicitly learns the shape and motion distribution of non-rigid objects through an iterative denoising process of compressed latent representations. The diffusion-based priors enable more plausible and probabilistic reconstructions when handling ambiguous inputs. We parameterize 4D dynamics with latent sets instead of using global latent codes. This novel 4D representation allows us to learn local shape and deformation patterns leading to more accurate non-linear motion capture and significantly improving generalizability to unseen motions and identities. For more temporally-coherent object tracking we synchronously denoise deformation latent sets and exchange information across multiple frames. To avoid computational overhead we designed an interleaved space and time attention block to alternately aggregate deformation latents along spatial and temporal domains. Extensive comparisons against state-of-the-art methods demonstrate the superiority of our Motion2VecSets in 4D reconstruction from various imperfect observations.
****
DiaLoc: An Iterative Approach to Embodied Dialog Localization
Chao Zhang, Mohan Li, Ignas Budvytis, Stephan Liwicki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12585-12593
Multimodal learning has advanced the performance for many vision-language tasks. However most existing works in embodied dialog research focus on navigation and leave the localization task understudied. The few existing dialog-based localization approaches assume the availability of entire dialog prior to localizaiton which is impractical for deployed dialog-based localization. In this paper we propose DiaLoc a new dialog-based localization framework which aligns with a real human operator behavior. Specifically we produce an iterative refinement of location predictions which can visualize current pose believes after each dialog turn. DiaLoc effectively utilizes the multimodal data for multi-shot localization where a fusion encoder fuses vision and dialog information iteratively. We achieve state-of-the-art results on embodied dialog-based localization task in single-shot (+7.08% in Acc5@valUnseen) and multi-shot settings (+10.85% in Acc5@valUnseen). DiaLoc narrows the gap between simulation and real-world applications opening doors for future research on collaborative localization and navigation.
****
Self-Training Large Language Models for Improved Visual Program Synthesis With Visual Reinforcement
Zaid Khan, Vijay Kumar BG, Samuel Schulter, Yun Fu, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14344-14353
Visual program synthesis is a promising approach to exploit the reasoning abilities of large language models for compositional computer vision tasks. Previous work has used few-shot prompting with frozen LLMs to synthesize visual programs. Training an LLM to write better visual programs is an attractive prospect but it is unclear how to accomplish this. No dataset of visual programs for training exists and acquisition of a visual program dataset cannot be easily crowdsourced due to the need for expert annotators. To get around the lack of direct supervision we explore improving the program synthesis abilities of an LLM using feedback from interactive experience. We propose a method where we exploit existing annotations for a vision-language task to improvise a coarse reward signal for that task treat the LLM as a policy and apply reinforced self-training to improve the visual program synthesis ability of the LLM for that task. We describe a series of experiments on object detection compositional visual question answering and image-text retrieval and show that in each case the self-trained LLM outperforms or performs on par with few-shot frozen LLMs that are an order of magnitude larger. Website: https://zaidkhan.me/ViReP
****
A2XP: Towards Private Domain Generalization
Geunhyeok Yu, Hyoseok Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23544-23553
Deep Neural Networks (DNNs) have become pivotal in various fields especially in

computer vision outperforming previous methodologies. A critical challenge in th
eir deployment is the bias inherent in data across different domains such as ima
ge style and environmental conditions leading to domain gaps. This necessitates
techniques for learning general representations from biased training data known
as domain generalization. This paper presents Attend to eXpert Prompts (A2XP) a
novel approach for domain generalization that preserves the privacy and integrit
y of the network architecture. A2XP consists of two phases: Expert Adaptation an
d Domain Generalization. In the first phase prompts for each source domain are o
ptimized to guide the model towards the optimal direction. In the second phase t
wo embedder networks are trained to effectively amalgamate these expert prompts
aiming for an optimal output. Our extensive experiments demonstrate that A2XP ac
hieves state-of-the-art results over existing non-private domain generalization
methods. The experimental results validate that the proposed approach not only t
ackles the domain generalization challenge in DNNs but also offers a privacy-pre
serving efficient solution to the broader field of computer vision.
****
Event-assisted Low-Light Video Object Segmentation
Hebei Li, Jin Wang, Jiahui Yuan, Yue Li, Wenming Weng, Yansong Peng, Yueyi Zhang
, Zhiwei Xiong, Xiaoyan Sun; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2024, pp. 3250-3259
In the realm of video object segmentation (VOS) the challenge of operating under
 low-light conditions persists resulting in notably degraded image quality and c
ompromised accuracy when comparing query and memory frames for similarity comput
ation. Event cameras characterized by their high dynamic range and ability to ca
pture motion information of objects offer promise in enhancing object visibility
 and aiding VOS methods under such low-light conditions. This paper introduces a
 pioneering framework tailored for low-light VOS leveraging event camera data to
 elevate segmentation accuracy. Our approach hinges on two pivotal components: t
he Adaptive Cross-Modal Fusion (ACMF) module aimed at extracting pertinent featu
res while fusing image and event modalities to mitigate noise interference and t
he Event-Guided Memory Matching (EGMM) module designed to rectify the issue of i
naccurate matching prevalent in low-light settings. Additionally we present the
creation of a synthetic LLE-DAVIS dataset and the curation of a real-world LLE-V
OS dataset encompassing frames and events. Experimental evaluations corroborate
the efficacy of our method across both datasets affirming its effectiveness in l
ow-light scenarios. The datasets are available at https://github.com/HebeiFast/E
ventLowLightVOS.
****
Active Domain Adaptation with False Negative Prediction for Object Detection
Yuzuru Nakamura, Yasunori Ishii, Takayoshi Yamashita; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28782-
28792
Domain adaptation adapts models to various scenes with different appearances. In
 this field active domain adaptation is crucial in effectively sampling a limite
d number of data in the target domain. We propose an active domain adaptation me
thod for object detection focusing on quantifying the undetectability of objects
. Existing methods for active sampling encounter challenges in considering undet
ected objects while estimating the uncertainty of model predictions. Our propose
d active sampling strategy addresses this issue using an active learning approac
h that simultaneously accounts for uncertainty and undetectability. Our newly pr
oposed False Negative Prediction Module evaluates the undetectability of images
containing undetected objects enabling more informed active sampling. This appro
ach considers previously overlooked undetected objects thereby reducing false ne
gative errors. Moreover using unlabeled data our proposed method utilizes uncert
ainty-guided pseudo-labeling to enhance domain adaptation further. Extensive exp
eriments demonstrate that the performance of our proposed method closely rivals
that of fully supervised learning while requiring only a fraction of the labelin
g efforts needed for the latter.
****
MLIP: Enhancing Medical Visual Representation with Divergence Encoder and Knowle

dge-guided Contrastive Learning
Zhe Li, Laurence T. Yang, Bocheng Ren, Xin Nie, Zhangyang Gao, Cheng Tan, Stan Z. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11704-11714
The scarcity of annotated data has sparked significant interest in unsupervised pre-training methods that leverage medical reports as auxiliary signals for medical visual representation learning. However existing research overlooks the multi-granularity nature of medical visual representation and lacks suitable contrastive learning techniques to improve the models' generalizability across different granularities leading to the underutilization of image-text information. To address this we propose MLIP a novel framework leveraging domain-specific medical knowledge as guiding signals to integrate language information into the visual domain through image-text contrastive learning. Our model includes global contrastive learning with our designed divergence encoder local token-knowledge-patch alignment contrastive learning and knowledge-guided category-level contrastive learning with expert knowledge. Experimental evaluations reveal the efficacy of our model in enhancing transfer performance for tasks such as image classification object detection and semantic segmentation. Notably MLIP surpasses state-of-the-art methods even with limited annotated data highlighting the potential of multimodal pre-training in advancing medical representation learning.
****
Generative 3D Part Assembly via Part-Whole-Hierarchy Message Passing
Bi'an Du, Xiang Gao, Wei Hu, Renjie Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20850-20859
Generative 3D part assembly involves understanding part relationships and predicting their 6-DoF poses for assembling a realistic 3D shape. Prior work often focus on the geometry of individual parts neglecting part-whole hierarchies of objects. Leveraging two key observations: 1) super-part poses provide strong hints about part poses and 2) predicting super-part poses is easier due to fewer super-parts we propose a part-whole-hierarchy message passing network for efficient 3D part assembly. We first introduce super-parts by grouping geometrically similar parts without any semantic labels. Then we employ a part-whole hierarchical encoder wherein a super-part encoder predicts latent super-part poses based on input parts. Subsequently we transform the point cloud using the latent poses feeding it to the part encoder for aggregating super-part information and reasoning about part relationships to predict all part poses. In training only ground-truth part poses are required. During inference the predicted latent poses of super-parts enhance interpretability. Experimental results on the PartNet dataset that our method achieves state-of-the-art performance in part and connectivity accuracy and enables an interpretable hierarchical part assembly.
****
VidToMe: Video Token Merging for Zero-Shot Video Editing
Xirui Li, Chao Ma, Xiaokang Yang, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7486-7495
Diffusion models have made significant advances in generating high-quality images but their application to video generation has remained challenging due to the complexity of temporal motion. Zero-shot video editing offers a solution by utilizing pre-trained image diffusion models to translate source videos into new ones. Nevertheless existing methods struggle to maintain strict temporal consistency and efficient memory consumption. In this work we propose a novel approach to enhance temporal consistency in generated videos by merging self-attention tokens across frames. By aligning and compressing temporally redundant tokens across frames our method improves temporal coherence and reduces memory consumption in self-attention computations. The merging strategy matches and aligns tokens according to the temporal correspondence between frames facilitating natural temporal consistency in generated video frames. To manage the complexity of video processing we divide videos into chunks and develop intra-chunk local token merging and inter-chunk global token merging ensuring both short-term video continuity and long-term content consistency. Our video editing approach seamlessly extends the advancements in image editing to video editing rendering favorable results in

temporal consistency over state-of-the-art methods.
****
FaceChain-SuDe: Building Derived Class to Inherit Category Attributes for One-shot Subject-Driven Generation
Pengchong Qiao, Lei Shang, Chang Liu, Baigui Sun, Xiangyang Ji, Jie Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7215-7224
Recently subject-driven generation has garnered significant interest due to its ability to personalize text-to-image generation. Typical works focus on learning the new subject's private attributes. However an important fact has not been taken seriously that a subject is not an isolated new concept but should be a specialization of a certain category in the pre-trained model. This results in the subject failing to comprehensively inherit the attributes in its category causing poor attribute-related generations. In this paper motivated by object-oriented programming we model the subject as a derived class whose base class is its semantic category. This modeling enables the subject to inherit public attributes from its category while learning its private attributes from the user-provided example. Specifically we propose a plug-and-play method Subject-Derived regularization (SuDe). It constructs the base-derived class modeling by constraining the subject-driven generated images to semantically belong to the subject's category. Extensive experiments under three baselines and two backbones on various subjects show that our SuDe enables imaginative attribute-related generations while maintaining subject fidelity. For the codes please refer to \href https://github.com/modelscope/facechain FaceChain .
****
Benchmarking Segmentation Models with Mask-Preserved Attribute Editing
Zijin Yin, Kongming Liang, Bing Li, Zhanyu Ma, Jun Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22509-22519
When deploying segmentation models in practice it is critical to evaluate their behaviors in varied and complex scenes. Different from the previous evaluation paradigms only in consideration of global attribute variations (e.g. adverse weather) we investigate both local and global attribute variations for robustness evaluation. To achieve this we construct a mask-preserved attribute editing pipeline to edit visual attributes of real images with precise control of structural information. Therefore the original segmentation labels can be reused for the edited images. Using our pipeline we construct a benchmark covering both object and image attributes (e.g. color material pattern style). We evaluate a broad variety of semantic segmentation models spanning from conventional close-set models to recent open-vocabulary large models on their robustness to different types of variations. We find that both local and global attribute variations affect segmentation performances and the sensitivity of models diverges across different variation types. We argue that local attributes have the same importance as global attributes and should be considered in the robustness evaluation of segmentation models. Code: https://github.com/PRIS-CV/Pascal-EA.
****
Analyzing and Improving the Training Dynamics of Diffusion Models
Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, Samuli Laine; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24174-24184
Diffusion models currently dominate the field of data-driven image synthesis with their unparalleled scaling to large datasets. In this paper we identify and rectify several causes for uneven and ineffective training in the popular ADM diffusion model architecture without altering its high-level structure. Observing uncontrolled magnitude changes and imbalances in both the network activations and weights over the course of training we redesign the network layers to preserve activation weight and update magnitudes on expectation. We find that systematic application of this philosophy eliminates the observed drifts and imbalances resulting in considerably better networks at equal computational complexity. Our modifications improve the previous record FID of 2.41 in ImageNet-512 synthesis to

1.81 achieved using fast deterministic sampling. As an independent contribution we present a method for setting the exponential moving average (EMA) parameters post-hoc i.e. after completing the training run. This allows precise tuning of EMA length without the cost of performing several training runs and reveals its surprising interactions with network architecture training time and guidance.
****
Hierarchical Correlation Clustering and Tree Preserving Embedding
Morteza Haghir Chehreghani, Mostafa Haghir Chehreghani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23083-23093
We propose a hierarchical correlation clustering method that extends the well-known correlation clustering to produce hierarchical clusters applicable to both positive and negative pairwise dissimilarities. Then in the following we study unsupervised representation learning with such hierarchical correlation clustering. For this purpose we first investigate embedding the respective hierarchy to be used for tree preserving embedding and feature extraction. Thereafter we study the extension of minimax distance measures to correlation clustering as another representation learning paradigm. Finally we demonstrate the performance of our methods on several datasets.
****
StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On
Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, Jaegul Choo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8176-8185
Given a clothing image and a person image an image-based virtual try-on aims to generate a customized image that appears natural and accurately reflects the characteristics of the clothing image. In this work we aim to expand the applicability of the pre-trained diffusion model so that it can be utilized independently for the virtual try-on task. The main challenge is to preserve the clothing details while effectively utilizing the robust generative capability of the pre-trained model. In order to tackle these issues we propose StableVITON learning the semantic correspondence between the clothing and the human body within the latent space of the pre-trained diffusion model in an end-to-end manner. Our proposed zero cross-attention blocks not only preserve the clothing details by learning the semantic correspondence but also generate high-fidelity images by utilizing the inherent knowledge of the pre-trained model in the warping process. Through our proposed novel attention total variation loss and applying augmentation we achieve the sharp attention map resulting in a more precise representation of clothing details. StableVITON outperforms the baselines in qualitative and quantitative evaluation showing promising quality in arbitrary person images. Our code is available at https://github.com/rlawjdghek/StableVITON.
****
Can Protective Perturbation Safeguard Personal Data from Being Exploited by Stable Diffusion?
Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, Xing Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24398-24407
Stable Diffusion has established itself as a foundation model in generative AI artistic applications receiving widespread research and application. Some recent fine-tuning methods have made it feasible for individuals to implant personalized concepts onto the basic Stable Diffusion model with minimal computational costs on small datasets. However these innovations have also given rise to issues like facial privacy forgery and artistic copyright infringement. In recent studies researchers have explored the addition of imperceptible adversarial perturbations to images to prevent potential unauthorized exploitation and infringements when personal data is used for fine-tuning Stable Diffusion. Although these studies have demonstrated the ability to protect images it is essential to consider that these methods may not be entirely applicable in real-world scenarios. In this paper we systematically evaluate the use of perturbations to protect images wit

hin a practical threat model. The results suggest that these approaches may not be sufficient to safeguard image privacy and copyright effectively. Furthermore we introduce a purification method capable of removing protected perturbations while preserving the original image structure to the greatest extent possible. Experiments reveal that Stable Diffusion can effectively learn from purified images over all protective methods.
****
Make-Your-Anchor: A Diffusion-based 2D Avatar Generation Framework
Ziyao Huang, Fan Tang, Yong Zhang, Xiaodong Cun, Juan Cao, Jintao Li, Tong-Yee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6997-7006
Despite the remarkable process of talking-head-based avatar-creating solutions directly generating anchor-style videos with full-body motions remains challenging. In this study we propose Make-Your-Anchor a novel system necessitating only a one-minute video clip of an individual for training subsequently enabling the automatic generation of anchor-style videos with precise torso and hand movements. Specifically we finetune a proposed structure-guided diffusion model on input video to render 3D mesh conditions into human appearances. We adopt a two-stage training strategy for the diffusion model effectively binding movements with specific appearances. To produce arbitrary long temporal video we extend the 2D U-Net in the frame-wise diffusion model to a 3D style without additional training cost and a simple yet effective batch-overlapped temporal denoising module is proposed to bypass the constraints on video length during inference. Finally a novel identity-specific face enhancement module is introduced to improve the visual quality of facial regions in the output videos. Comparative experiments demonstrate the effectiveness and superiority of the system in terms of visual quality temporal coherence and identity preservation outperforming SOTA diffusion/non-diffusion methods. Project page: https://github.com/ICTMCG/Make-Your-Anchor.
****
MultiPLY: A Multisensory Object-Centric Embodied Large Language Model in 3D World
Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26406-26416
Human beings possess the capability to multiply a melange of multisensory cues while actively exploring and interacting with the 3D world. Current multi-modal large language models however passively absorb sensory data as inputs lacking the capacity to actively interact with the objects in the 3D environment and dynamically collect their multisensory information. To usher in the study of this area we propose MultiPLY a multisensory embodied large language model that could incorporate multisensory interactive data including visual audio tactile and thermal information into large language models thereby establishing the correlation among words actions and percepts. To this end we first collect Multisensory Universe a large-scale multisensory interaction dataset comprising 500k data by deploying an LLM-powered embodied agent to engage with the 3D environment. To perform instruction tuning with pre-trained LLM on such generated data we first encode the 3D scene as abstracted object-centric representations and then introduce action tokens denoting that the embodied agent takes certain actions within the environment as well as state tokens that represent the multisensory state observations of the agent at each time step. In the inference time MultiPLY could generate action tokens instructing the agent to take the action in the environment and obtain the next multisensory state observation. The observation is then appended back to the LLM via state tokens to generate subsequent text or action tokens. We demonstrate that MultiPLY outperforms baselines by a large margin through a diverse set of embodied tasks involving object retrieval tool use multisensory captioning and task decomposition.
****
Learning to Visually Localize Sound Sources from Mixtures without Prior Source Knowledge
Dongjin Kim, Sung Jin Um, Sangmin Lee, Jung Uk Kim; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26467-26476
The goal of the multi-sound source localization task is to localize sound sources from the mixture individually. While recent multi-sound source localization methods have shown improved performance they face challenges due to their reliance on prior information about the number of objects to be separated. In this paper to overcome this limitation we present a novel multi-sound source localization method that can perform localization without prior knowledge of the number of sound sources. To achieve this goal we propose an iterative object identification (IOI) module which can recognize sound-making objects in an iterative manner. After finding the regions of sound-making objects we devise object similarity-aware clustering (OSC) loss to guide the IOI module to effectively combine regions of the same object but also distinguish between different objects and backgrounds. It enables our method to perform accurate localization of sound-making objects without any prior knowledge. Extensive experimental results on the MUSIC and VGGSound benchmarks show the significant performance improvements of the proposed method over the existing methods for both single and multi-source. Our code is available at: https://github.com/VisualAIKHU/NoPrior_MultiSSL
****
Learning Dynamic Tetrahedra for High-Quality Talking Head Synthesis
Zicheng Zhang, Ruobing Zheng, Bonan Li, Congying Han, Tianqi Li, Meng Wang, Tiande Guo, Jingdong Chen, Ziwen Liu, Ming Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5209-5219
Recent works in implicit representations such as Neural Radiance Fields (NeRF) have advanced the generation of realistic and animatable head avatars from video sequences. These implicit methods are still confronted by visual artifacts and jitters since the lack of explicit geometric constraints poses a fundamental challenge in accurately modeling complex facial deformations. In this paper we introduce Dynamic Tetrahedra (DynTet) a novel hybrid representation that encodes explicit dynamic meshes by neural networks to ensure geometric consistency across various motions and viewpoints. DynTet is parameterized by the coordinate-based networks which learn signed distance deformation and material texture anchoring the training data into a predefined tetrahedra grid. Leveraging Marching Tetrahedra DynTet efficiently decodes textured meshes with a consistent topology enabling fast rendering through a differentiable rasterizer and supervision via a pixel loss. To enhance training efficiency we incorporate classical 3D Morphable Models to facilitate geometry learning and define a canonical space for simplifying texture learning. These advantages are readily achievable owing to the effective geometric representation employed in DynTet. Compared with prior works DynTet demonstrates significant improvements in fidelity lip synchronization and real-time performance according to various metrics. Beyond producing stable and visually appealing synthesis videos our method also outputs the dynamic meshes which is promising to enable many emerging applications. Code is available at https://github.com/zhangzc21/DynTet.
****
Collaborative Learning of Anomalies with Privacy (CLAP) for Unsupervised Video Anomaly Detection: A New Baseline
Anas Al-lahham, Muhammad Zaigham Zaheer, Nurbek Tastan, Karthik Nandakumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12416-12425
nsupervised (US) video anomaly detection (VAD) in surveillance applications is gaining more popularity lately due to its practical real-world applications. Due to the extremely challenging nature of this task where learning is carried out without any annotations privacy-critical collaborative learning of US-VAD systems has not been studied yet. As surveillance videos are privacy sensitive and the availability of large-scale video data may enable better US-VAD systems collaborative learning can be highly rewarding in this setting. In this paper we propose a new baseline for anomaly detection capable of localizing anomalous events in complex surveillance scenarios in a fully unsupervised fashion without any labels on a privacy-retaining participant-based distributed training configuration. A

dditionally we propose three new evaluation protocols to extensively evaluate an omaly detection approaches on various scenarios of collaborations and data avail ability. Moreover based on these protocols we modify existing VAD datasets to ex tensively evaluate our approach as well as existing US SOTA methods on two large -scale datasets including UCF-Crime and XD-Violence. All proposed evaluation pro tocols dataset splits and codes are available here: \href https://github.com/Ana sEmad11/CLAP https://github.com/AnasEmad11/CLAP .
****
Regressor-Segmenter Mutual Prompt Learning for Crowd Counting
Mingyue Guo, Li Yuan, Zhaoyi Yan, Binghui Chen, Yaowei Wang, Qixiang Ye; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28380-28389
Crowd counting has achieved significant progress by training regressors to predi ct instance positions. In heavily crowded scenarios however regressors are chall enged by uncontrollable annotation variance which causes density map bias and co ntext information inaccuracy. In this study we propose mutual prompt learning (m Prompt) which leverages a regressor and a segmenter as guidance for each other s olving bias and inaccuracy caused by annotation variance while distinguishing fo reground from background. In specific mPrompt leverages point annotations to tun e the segmenter and predict pseudo head masks in a way of point prompt learning. It then uses the predicted segmentation masks which serve as spatial constraint to rectify biased point annotations as context prompt learning. mPrompt defines a way of mutual information maximization from prompt learning mitigating the im pact of annotation variance while improving model accuracy. Experiments show tha t mPrompt significantly reduces the Mean Average Error (MAE) demonstrating the p otential to be general framework for down-stream vision tasks. Code is available at https://github.com/csguomy/mPrompt.
****
Instantaneous Perception of Moving Objects in 3D
Di Liu, Bingbing Zhuang, Dimitris N. Metaxas, Manmohan Chandraker; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 19573-19583
The perception of 3D motion of surrounding traffic participants is crucial for d riving safety. While existing works primarily focus on general large motions we contend that the instantaneous detection and quantification of subtle motions is equally important as they indicate the nuances in driving behavior that may be safety critical such as behaviors near a stop sign of parking positions. We delv e into this under-explored task examining its unique challenges and developing o ur solution accompanied by a carefully designed benchmark. Specifically due to t he lack of correspondences between consecutive frames of sparse Lidar point clou ds static objects might appear to be moving - the so-called swimming effect. Thi s intertwines with the true object motion thereby posing ambiguity in accurate e stimation especially for subtle motion. To address this we propose to leverage l ocal occupancy completion of object point clouds to densify the shape cue and mi tigate the impact of swimming artifacts. The occupancy completion is learned in an end-to-end fashion together with the detection of moving objects and the esti mation of their motion instantaneously as soon as objects start to move. Extensi ve experiments demonstrate superior performance compared to standard 3D motion e stimation approaches particularly highlighting our method's specialized treatmen t of subtle motion.
****
CORE-MPI: Consistency Object Removal with Embedding MultiPlane Image
Donggeun Yoon, Donghyeon Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20081-20090
Novel view synthesis is attractive for social media but it often contains unwant ed details such as personal information that needs to be edited out for a better experience. Multiplane image (MPI) is desirable for social media because of its generality but it is complex and computationally expensive making object remova l challenging. To address these challenges we propose CORE-MPI which employs emb edding images to improve the consistency and accessibility of MPI object removal

. CORE-MPI allows for real-time transmission and interaction with embedding imag
es on social media facilitating object removal with a single mask. However recov
ering the geometric information hidden in the embedding images is a significant
challenge. Therefore we propose a dual-network approach where one network focuse
s on color restoration and the other on inpainting the embedding image including
 geometric information. For the training of CORE-MPI we introduce a pseudo-refer
ence loss aimed at proficient color recovery even in complex scenes or with larg
e masks. Furthermore we present a disparity consistency loss to preserve the geo
metric consistency of the inpainted region. We demonstrate the effectiveness of
CORE-MPI on RealEstate10K and UCSD datasets.
****
3D Geometry-Aware Deformable Gaussian Splatting for Dynamic View Synthesis
Zhicheng Lu, Xiang Guo, Le Hui, Tianrui Chen, Min Yang, Xiao Tang, Feng Zhu, Yuc
hao Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2024, pp. 8900-8910
In this paper we propose a 3D geometry-aware deformable Gaussian Splatting metho
d for dynamic view synthesis. Existing neural radiance fields (NeRF) based solut
ions learn the deformation in an implicit manner which cannot incorporate 3D sce
ne geometry. Therefore the learned deformation is not necessarily geometrically
coherent which results in unsatisfactory dynamic view synthesis and 3D dynamic r
econstruction. Recently 3D Gaussian Splatting provides a new representation of t
he 3D scene building upon which the 3D geometry could be exploited in learning t
he complex 3D deformation. Specifically the scenes are represented as a collecti
on of 3D Gaussian where each 3D Gaussian is optimized to move and rotate over ti
me to model the deformation. To enforce the 3D scene geometry constraint during
deformation we explicitly extract 3D geometry features and integrate them in lea
rning the 3D deformation. In this way our solution achieves 3D geometry-aware de
formation modeling which enables improved dynamic view synthesis and 3D dynamic
reconstruction. Extensive experimental results on both synthetic and real datase
ts prove the superiority of our solution which achieves new state-of-the-art per
formance. The project is available at \href https://npucvr.github.io/GaGS/ http
s://npucvr.github.io/GaGS/ .
****
Person-in-WiFi 3D: End-to-End Multi-Person 3D Pose Estimation with Wi-Fi
Kangwei Yan, Fei Wang, Bo Qian, Han Ding, Jinsong Han, Xing Wei; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 969-978
Wi-Fi signals in contrast to cameras offer privacy protection and occlusion resi
lience for some practical scenarios such as smart homes elderly care and virtual
 reality. Recent years have seen remarkable progress in the estimation of single
-person 2D pose single-person 3D pose and multi-person 2D pose. This paper takes
 a step forward by introducing Person-in-WiFi 3D a pioneering Wi-Fi system that
accomplishes multi-person 3D pose estimation. Person-in-WiFi 3D has two main upd
ates. Firstly it has a greater number of Wi-Fi devices to enhance the capability
 for capturing spatial reflections from multiple individuals. Secondly it levera
ges the Transformer for end-to-end estimation. Compared to its predecessor Perso
n-in-WiFi 3D is storage-efficient and fast. We deployed a proof-of-concept syste
m in 4mx3.5m areas and collected a dataset of over 97K frames with seven volunte
ers. Person-in-WiFi 3D attains 3D joint localization errors of 91.7mm (1-person)
 108.1mm (2-person) and 125.3mm (3-person) comparable to cameras and millimeter-
wave radars.
****
Backpropagation-free Network for 3D Test-time Adaptation
Yanshuo Wang, Ali Cheraghian, Zeeshan Hayder, Jie Hong, Sameera Ramasinghe, Shaf
in Rahman, David Ahmedt-Aristizabal, Xuesong Li, Lars Petersson, Mehrtash Harand
i; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2024, pp. 23231-23241
Real-world systems often encounter new data over time which leads to experiencin
g target domain shifts. Existing Test-Time Adaptation (TTA) methods tend to appl
y computationally heavy and memory-intensive backpropagation-based approaches to

handle this. Here we propose a novel method that uses a backpropagation-free ap proach for TTA for the specific case of 3D data. Our model uses a two-stream arc hitecture to maintain knowledge about the source domain as well as complementary target-domain-specific information. The backpropagation-free property of our mo del helps address the well-known forgetting problem and mitigates the error accu mulation issue. The proposed method also eliminates the need for the usually noi sy process of pseudo-labeling and reliance on costly self-supervised training. M oreover our method leverages subspace learning effectively reducing the distribu tion variance between the two domains. Furthermore the source-domain-specific an d the target-domain-specific streams are aligned using a novel entropy-based ada ptive fusion strategy. Extensive experiments on popular benchmarks demonstrate t he effectiveness of our method. The code will be available at https://github.com /abie-e/BFTT3D.
****
Resource-Efficient Transformer Pruning for Finetuning of Large Models
Fatih Ilhan, Gong Su, Selim Furkan Tekin, Tiansheng Huang, Sihao Hu, Ling Liu; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16206-16215
With the recent advances in vision transformers and large language models (LLMs) finetuning costly large models on downstream learning tasks poses significant c hallenges under limited computational resources. This paper presents a REsource and ComputAtion-efficient Pruning framework (RECAP) for the finetuning of transf ormer-based large models. RECAP by design bridges the gap between efficiency and performance through an iterative process cycling between pruning finetuning and updating stages to explore different chunks of the given large-scale model. At each iteration we first prune the model with Taylor-approximation-based importan ce estimation and then only update a subset of the pruned model weights based on the Fisher-information criterion. In this way RECAP achieves two synergistic an d yet conflicting goals: reducing the GPU memory footprint while maintaining mod el performance unlike most existing pruning methods that require the model to be finetuned beforehand for better preservation of model performance. We perform e xtensive experiments with a wide range of large transformer-based architectures on various computer vision and natural language understanding tasks. Compared to recent pruning techniques we demonstrate that RECAP offers significant improvem ents in GPU memory efficiency capable of reducing the footprint by up to 65%.
****
ParamISP: Learned Forward and Inverse ISPs using Camera Parameters
Woohyeok Kim, Geonu Kim, Junyong Lee, Seungyong Lee, Seung-Hwan Baek, Sunghyun C ho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2024, pp. 26067-26076
RAW images are rarely shared mainly due to its excessive data size compared to t heir sRGB counterparts obtained by camera ISPs. Learning the forward and inverse processes of camera ISPs has been recently demonstrated enabling physically-mea ningful RAW-level image processing on input sRGB images. However existing learni ng-based ISP methods fail to handle the large variations in the ISP processes wi th respect to camera parameters such as ISO and exposure time and have limitatio ns when used for various applications. In this paper we propose ParamISP a learn ing-based method for forward and inverse conversion between sRGB and RAW images that adopts a novel neural-network module to utilize camera parameters which is dubbed as ParamNet. Given the camera parameters provided in the EXIF data ParamN et converts them into a feature vector to control the ISP networks. Extensive ex periments demonstrate that ParamISP achieve superior RAW and sRGB reconstruction results compared to previous methods and it can be effectively used for a varie ty of applications such as deblurring dataset synthesis raw deblurring HDR recon struction and camera-to-camera transfer.
****
Perturbing Attention Gives You More Bang for the Buck: Subtle Imaging Perturbati ons That Efficiently Fool Customized Diffusion Models
Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, Xiang Wei; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR

), 2024, pp. 24534-24543
Diffusion models (DMs) embark a new era of generative modeling and offer more op portunities for efficient generating high-quality and realistic data samples. Ho wever their widespread use has also brought forth new challenges in model securi ty which motivates the creation of more effective adversarial attackers on DMs t o understand its vulnerability. We propose CAAT a simple but generic and efficie nt approach that does not require costly training to effectively fool latent dif fusion models (LDMs). The approach is based on the observation that cross-attent ion layers exhibits higher sensitivity to gradient change allowing for leveragin g subtle perturbations on published images to significantly corrupt the generate d images. We show that a subtle perturbation on an image can significantly impac t the cross-attention layers thus changing the mapping between text and image du ring the fine-tuning of customized diffusion models. Extensive experiments demon strate that CAAT is compatible with diverse diffusion models and outperforms bas eline attack methods in a more effective (more noise) and efficient (twice as fa st as Anti-DreamBooth and Mist) manner.
****
Fairy: Fast Parallelized Instruction-Guided Video-to-Video Synthesis
Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, Peter Vajda; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8261-8270
In this paper we introduce Fairy a minimalist yet robust adaptation of image-edi ting diffusion models enhancing them for video editing applications. Our approac h centers on the concept of anchor-based cross-frame attention a mechanism that implicitly propagates diffusion features across frames ensuring superior tempora l coherence and high-fidelity synthesis. Fairy not only addresses limitations of previous models including memory and processing speed. It also improves tempora l consistency through a unique data augmentation strategy. This strategy renders the model equivariant to affine transformations in both source and target image s. Remarkably efficient Fairy generates 120-frame 512x384 videos (4-second durat ion at 30 FPS) in just 14 seconds outpacing prior works by at least 44x. A compr ehensive user study involving 1000 generated samples confirms that our approach delivers superior quality decisively outperforming established methods.
****
SmartEdit: Exploring Complex Instruction-based Image Editing with Multimodal Lar ge Language Models
Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, J iantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, Ying Shan; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8362-8371
Current instruction-based image editing methods such as InstructPix2Pix often fa il to produce satisfactory results in complex scenarios due to their dependence on the simple CLIP text encoder in diffusion models. To rectify this this paper introduces SmartEdit a novel approach of instruction-based image editing that le verages Multimodal Large Language Models (MLLMs) to enhance its understanding an d reasoning capabilities. However direct integration of these elements still fac es challenges in situations requiring complex reasoning. To mitigate this we pro pose a Bidirectional Interaction Module (BIM) that enables comprehensive bidirec tional information interactions between the input image and the MLLM output. Dur ing training we initially incorporate perception data to boost the perception an d understanding capabilities of diffusion models. Subsequently we demonstrate th at a small amount of complex instruction editing data can effectively stimulate SmartEdit's editing capabilities for more complex instructions. We further const ruct a new evaluation dataset Reason-Edit specifically tailored for complex inst ruction-based image editing. Both quantitative and qualitative results on this e valuation dataset indicate that our SmartEdit surpasses previous methods paving the way for the practical application of complex instruction-based image editing .
****
SeNM-VAE: Semi-Supervised Noise Modeling with Hierarchical Variational Autoencod

er
Dihan Zheng, Yihang Zou, Xiaowen Zhang, Chenglong Bao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25889-25899
The data bottleneck has emerged as a fundamental challenge in learning based image restoration methods. Researchers have attempted to generate synthesized training data using paired or unpaired samples to address this challenge. This study proposes SeNM-VAE a semi-supervised noise modeling method that leverages both paired and unpaired datasets to generate realistic degraded data. Our approach is based on modeling the conditional distribution of degraded and clean images with a specially designed graphical model. Under the variational inference framework we develop an objective function for handling both paired and unpaired data. We employ our method to generate paired training samples for real-world image denoising and super-resolution tasks. Our approach excels in the quality of synthetic degraded images compared to other unpaired and paired noise modeling methods. Furthermore our approach demonstrates remarkable performance in downstream image restoration tasks even with limited paired data. With more paired data our method achieves the best performance on the SIDD dataset.
****
Multimodal Industrial Anomaly Detection by Crossmodal Feature Mapping
Alex Costanzino, Pierluigi Zama Ramirez, Giuseppe Lisanti, Luigi Di Stefano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17234-17243
Recent advancements have shown the potential of leveraging both point clouds and images to localize anomalies. Nevertheless their applicability in industrial manufacturing is often constrained by significant drawbacks such as the use of memory banks which leads to a substantial increase in terms of memory footprint and inference times. We propose a novel light and fast framework that learns to map features from one modality to the other on nominal samples and detect anomalies by pinpointing inconsistencies between observed and mapped features. Extensive experiments show that our approach achieves state-of-the-art detection and segmentation performance in both the standard and few-shot settings on the MVTec 3D-AD dataset while achieving faster inference and occupying less memory than previous multimodal AD methods. Furthermore we propose a layer pruning technique to improve memory and time efficiency with a marginal sacrifice in performance.
****
FFF: Fixing Flawed Foundations in Contrastive Pre-Training Results in Very Strong Vision-Language Models
Adrian Bulat, Yassine Ouali, Georgios Tzimiropoulos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14172-14182
Despite noise and caption quality having been acknowledged as important factors impacting vision-language contrastive pre-training in this paper we show that the full potential of improving the training process by addressing such issues is yet to be realized. Specifically we firstly study and analyze two issues affecting training: incorrect assignment of negative pairs and low caption quality and diversity. Then we devise effective solutions for addressing both problems which essentially require training with multiple true positive pairs. Finally we propose training with sigmoid loss to address such a requirement. We show very large gains over the current state-of-the-art for both image recognition (+6% on average over 11 datasets) and image retrieval (+19% on Flickr30k and +15% on MSCOCO).
****
Anchor-based Robust Finetuning of Vision-Language Models
Jinwei Han, Zhiwen Lin, Zhongyisun Sun, Yingguo Gao, Ke Yan, Shouhong Ding, Yuan Gao, Gui-Song Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26919-26928
We aim at finetuning a vision-language model without hurting its out-of-distribution (OOD) generalization. We address two types of OOD generalization i.e. i) domain shift such as natural to sketch images and ii) zero-shot capability to reco

gnize the category that was not contained in the finetune data. Arguably the dim inished OOD generalization after finetuning stems from the excessively simplifie d finetuning target which only provides the class information such as "a photo o f a [CLASS]". This is distinct from the process in that CLIP was pretrained wher e there is abundant text supervision with rich semantic information. Therefore w e propose to compensate for the finetune process using auxiliary supervision wit h rich semantic information which acts as anchors to preserve the OOD generaliza tion. Specifically two types of anchors are elaborated in our methods including i) text-compensated anchor which uses the images from the finetune set but enric hes the text supervision from a pretrained captioner ii) image-text-pair anchor which is retrieved from the dataset similar to pretraining data of CLIP accordin g to the downstream task associating with the original CLIP text with rich seman tics. Those anchors are utilized as auxiliary semantic information to maintain t he original feature space of CLIP thereby preserving the OOD generalization capa bilities. Comprehensive experiments demonstrate that our method achieves in-dist ribution performance akin to conventional finetuning while attaining new state-o f-the-art results on domain shift and zero-shot learning benchmarks.
****
Low-power Continuous Remote Behavioral Localization with Event Cameras
Friedhelm Hamann, Suman Ghosh, Ignacio Juarez Martinez, Tom Hart, Alex Kacelnik, Guillermo Gallego; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2024, pp. 18612-18621
Researchers in natural science need reliable methods for quantifying animal beha vior. Recently numerous computer vision methods emerged to automate the process. However observing wild species at remote locations remains a challenging task d ue to difficult lighting conditions and constraints on power supply and data sto rage. Event cameras offer unique advantages for battery-dependent remote monitor ing due to their low power consumption and high dynamic range capabilities. We u se this novel sensor to quantify a behavior in Chinstrap penguins called ecstati c display. We formulate the problem as a temporal action detection task determin ing the start and end times of the behavior. For this purpose we recorded a colo ny of breeding penguins in Antarctica for several weeks and labeled event data o n 16 nests. The developed method consists of a generator of candidate time inter vals (proposals) and a classifier of the actions within them. The experiments sh ow that the event cameras' natural response to motion is effective for continuou s behavior monitoring and detection reaching a mean average precision (mAP) of 5 8% (which increases to 63% in good weather conditions). The results also demonst rate the robustness against various lighting conditions contained in the challen ging dataset. The low-power capabilities of the event camera allow it to record significantly longer than with a conventional camera. This work pioneers the use of event cameras for remote wildlife observation opening new interdisciplinary opportunities. https:// tub-rip.github.io/ eventpenguins/
****
SportsHHI: A Dataset for Human-Human Interaction Detection in Sports Videos
Tao Wu, Runyu He, Gangshan Wu, Limin Wang; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18537-18546
Video-based visual relation detection tasks such as video scene graph generation play important roles in fine-grained video understanding. However current video visual relation detection datasets have two main limitations that hinder the pr ogress of research in this area. First they do not explore complex human-human i nteractions in multi-person scenarios. Second the relation types of existing dat asets have relatively low-level semantics and can be often recognized by appeara nce or simple prior information without the need for detailed spatio-temporal co ntext reasoning. Nevertheless comprehending high-level interactions between huma ns is crucial for understanding complex multi-person videos such as sports and s urveillance videos. To address this issue we propose a new video visual relation detection task: video human-human interaction detection and build a dataset nam ed SportsHHI for it. SportsHHI contains 34 high-level interaction classes from b asketball and volleyball sports. 118075 human bounding boxes and 50649 interacti on instances are annotated on 11398 keyframes. To benchmark this we propose a tw

o-stage baseline method and conduct extensive experiments to reveal the key fact ors for a successful human-human interaction detector. We hope that SportsHHI ca n stimulate research on human interaction understanding in videos and promote th e development of spatio-temporal context modeling techniques in video visual rel ation detection.
****
DiSR-NeRF: Diffusion-Guided View-Consistent Super-Resolution NeRF
Jie Long Lee, Chen Li, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2024, pp. 20561-20570
We present DiSR-NeRF a diffusion-guided framework for view-consistent super-reso lution (SR) NeRF. Unlike prior works we circumvent the requirement for high-reso lution (HR) reference images by leveraging existing powerful 2D super-resolution models. Nonetheless independent SR 2D images are often inconsistent across diff erent views. We thus propose Iterative 3D Synchronization (I3DS) to mitigate the inconsistency problem via the inherent multi-view consistency property of NeRF. Specifically our I3DS alternates between upscaling low-resolution (LR) rendered images with diffusion models and updating the underlying 3D representation with standard NeRF training. We further introduce Renoised Score Distillation (RSD) a novel score-distillation objective for 2D image resolution. Our RSD combines f eatures from ancestral sampling and Score Distillation Sampling (SDS) to generat e sharp images that are also LR-consistent. Qualitative and quantitative results on both synthetic and real-world datasets demonstrate that our DiSR-NeRF can ac hieve better results on NeRF super-resolution compared with existing works. Code and video results available at the project website.
****
Dispersed Structured Light for Hyperspectral 3D Imaging
Suhyun Shin, Seokjun Choi, Felix Heide, Seung-Hwan Baek; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 249 97-25006
Hyperspectral 3D imaging aims to acquire both depth and spectral information of a scene. However existing methods are either prohibitively expensive and bulky o r compromise on spectral and depth accuracy. In this paper we present Dispersed Structured Light (DSL) a cost-effective and compact method for accurate hyperspe ctral 3D imaging. DSL modifies a traditional projector-camera system by placing a sub-millimeter thick diffraction grating film front of the projector. This con figuration enables dispersing structured light based on light wavelength. To uti lize the dispersed structured light we devise a model for dispersive projection image formation and a per-pixel hyperspectral 3D reconstruction method. We valid ate DSL by instantiating a compact experimental prototype. DSL achieves spectral accuracy of 18.8nm full-width half-maximum (FWHM) and depth error of 1mm outper forming prior work on practical hyperspectral 3D imaging. DSL promises accurate and practical hyperspectral 3D imaging for diverse application domains including computer vision and graphics cultural heritage geology and biology.
****
CrowdDiff: Multi-hypothesis Crowd Density Estimation using Diffusion Models
Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, Vis hal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2024, pp. 12809-12819
Crowd counting is a fundamental problem in crowd analysis which is typically acc omplished by estimating a crowd density map and summing over the density values. However this approach suffers from background noise accumulation and loss of de nsity due to the use of broad Gaussian kernels to create the ground truth densit y maps. This issue can be overcome by narrowing the Gaussian kernel. However exi sting approaches perform poorly when trained with ground truth density maps with broad kernels. To deal with this limitation we propose using conditional diffus ion models to predict density maps as diffusion models show high fidelity to tra ining data during generation. With that we present CrowdDiff that generates the crowd density map as a reverse diffusion process. Furthermore as the intermediat e time steps of the diffusion process are noisy we incorporate a regression bran ch for direct crowd estimation only during training to improve the feature learn

ing. In addition owing to the stochastic nature of the diffusion model we introd uce producing multiple density maps to improve the counting performance contrary to the existing crowd counting pipelines. We conduct extensive experiments on p ublicly available datasets to validate the effectiveness of our method. CrowdDif f outperforms existing \sota crowd counting methods on several public crowd anal ysis benchmarks with significant improvements. CrowdDiff project is available at : https://dylran.github.io/crowddiff.github.io.
****
It's All About Your Sketch: Democratising Sketch Control in Diffusion Models
Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nat h Chowdhury, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2024, pp. 7204-7214
This paper unravels the potential of sketches for diffusion models addressing th e deceptive promise of direct sketch control in generative AI. We importantly de mocratise the process enabling amateur sketches to generate precise images livin g up to the commitment of "what you sketch is what you get". A pilot study under scores the necessity revealing that deformities in existing models stem from spa tial-conditioning. To rectify this we propose an abstraction-aware framework uti lising a sketch adapter adaptive time-step sampling and discriminative guidance from a pre-trained fine-grained sketch-based image retrieval model working syner gistically to reinforce fine-grained sketch-photo association. Our approach oper ates seamlessly during inference without the need for textual prompts; a simple rough sketch akin to what you and I can create suffices! We welcome everyone to examine results presented in the paper and its supplementary. Contributions incl ude democratising sketch control introducing an abstraction-aware framework and leveraging discriminative guidance validated through extensive experiments.
****
GLID: Pre-training a Generalist Encoder-Decoder Vision Model
Jihao Liu, Jinliang Zheng, Yu Liu, Hongsheng Li; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22851-22860
This paper proposes a GeneraLIst encoder-Decoder (GLID) pre-training method for better handling various downstream computer vision tasks. While self-supervised pre-training approaches e.g. Masked Autoencoder have shown success in transfer l earning task-specific sub-architectures are still required to be appended for di fferent downstream tasks which cannot enjoy the benefits of large-scale pre-trai ning. GLID overcomes this challenge by allowing the pre-trained generalist encod er-decoder to be fine-tuned on various vision tasks with minimal task-specific a rchitecture modifications. In the GLID training scheme pre-training pretext task and other downstream tasks are modeled as "query-to-answer" problems including the pre-training pretext task and other downstream tasks. We pre-train a task-ag nostic encoder-decoder with query-mask pairs. During fine-tuning GLID maintains the pre-trained encoder-decoder and queries only replacing the topmost linear tr ansformation layer with task-specific linear heads. This minimizes the pretrain-finetune architecture inconsistency and enables the pre-trained model to better adapt to downstream tasks. GLID achieves competitive performance on various visi on tasks including object detection image segmentation pose estimation and depth estimation outperforming or matching specialist models such as Mask2Former DETR ViTPose and BinsFormer.
****
Diffusion-FOF: Single-View Clothed Human Reconstruction via Diffusion-Based Four ier Occupancy Field
Yuanzhen Li, Fei Luo, Chunxia Xiao; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2024, pp. 9525-9534
Reconstructing a clothed human from a single-view image has several challenging issues including flexibly representing various body shapes and poses estimating complete 3D geometry and consistent texture and achieving more fine-grained deta ils. To address them we propose a new diffusion-based Fourier occupancy field me thod to improve the human representing ability and the geometry generating abili ty. First we estimate the back-view image from the given reference image by inco rporating a style consistency constraint. Then we extract multi-scale features o

f the two images as conditional and design a diffusion model to generate the Fou
rier occupancy field in the wavelet domain. We refine the initial estimated Four
ier occupancy field with image features as conditions to improve the geometric a
ccuracy. Finally the reference and estimated back-view images are mapped onto th
e human model creating a textured clothed human model. Substantial experiments a
re conducted and the experimental results show that our method outperforms the s
tate-of-the-art methods in geometry and texture reconstruction performance.
****
When StyleGAN Meets Stable Diffusion: a W+ Adapter for Personalized Image Genera
tion
Xiaoming Li, Xinyu Hou, Chen Change Loy; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2187-2196
Text-to-image diffusion models have remarkably excelled in producing diverse hig
h-quality and photo-realistic images. This advancement has spurred a growing int
erest in incorporating specific identities into generated content. Most current
methods employ an inversion approach to embed a target visual concept into the t
ext embedding space using a single reference image. However the newly synthesize
d faces either closely resemble the reference image in terms of facial attribute
s such as expression or exhibit a reduced capacity for identity preservation. Te
xt descriptions intended to guide the facial attributes of the synthesized face
may fall short owing to the intricate entanglement of identity information with
identity-irrelevant facial attributes derived from the reference image. To addre
ss these issues we present the novel use of the extended StyleGAN embedding spac
e \mathcal W _+ to achieve enhanced identity preservation and disentanglement fo
r diffusion models. By aligning this semantically meaningful human face latent s
pace with text-to-image diffusion models we succeed in maintaining high fidelity
in identity preservation coupled with the capacity for semantic editing. Additi
onally we propose new training objectives to balance the influences of both prom
pt and identity conditions ensuring that the identity-irrelevant background rema
ins \lxm negligibly affected during facial attribute modifications. Extensive e
xperiments reveal that our method adeptly generates personalized text-to-image o
utputs that are not only compatible with prompt descriptions but also amenable t
o common StyleGAN editing directions in diverse settings. Our code and model are
available at https://github.com/csxmli2016/w-plus-adapter.
****
ToNNO: Tomographic Reconstruction of a Neural Network's Output for Weakly Superv
ised Segmentation of 3D Medical Images
Marius Schmidt-Mengin, Alexis Benichoux, Shibeshih Belachew, Nikos Komodakis, Ni
kos Paragios; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2024, pp. 11428-11438
Annotating lots of 3D medical images for training segmentation models is time-co
nsuming. The goal of weakly supervised semantic segmentation is to train segment
ation models without using any ground truth segmentation masks. Our work address
es the case where only image-level categorical labels indicating the presence or
absence of a particular region of interest (such as tumours or lesions) are ava
ilable. Most existing methods rely on class activation mapping (CAM). We propose
a novel approach ToNNO which is based on the Tomographic reconstruction of a Ne
ural Network's Output. Our technique extracts stacks of slices with different an
gles from the input 3D volume feeds these slices to a 2D encoder and applies the
inverse Radon transform in order to reconstruct a 3D heatmap of the encoder's p
redictions. This generic method allows to perform dense prediction tasks on 3D v
olumes using any 2D image encoder. We apply it to weakly supervised medical imag
e segmentation by training the 2D encoder to output high values for slices conta
ining the regions of interest. We test it on four large scale medical image data
sets and outperform 2D CAM methods. We then extend ToNNO by combining tomographi
c reconstruction with CAM methods proposing Averaged CAM and Tomographic CAM whi
ch obtain even better results.
****
Learning to Navigate Efficiently and Precisely in Real Environments
Guillaume Bono, Hervé Poirier, Leonid Antsfeld, Gianluca Monaci, Boris Chidlovsk

ii, Christian Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17837-17846

In the context of autonomous navigation of terrestrial robots the creation of realistic models for agent dynamics and sensing is a widespread habit in the robotics literature and in commercial applications where they are used for model based control and/or for localization and mapping. The more recent Embodied AI literature on the other hand focuses on modular or end-to-end agents trained in simulators like Habitat or AI-Thor where the emphasis is put on photo-realistic rendering and scene diversity but high-fidelity robot motion is assigned a less privileged role. The resulting sim2real gap significantly impacts transfer of the trained models to real robotic platforms. In this work we explore end-to-end training of agents in simulation in settings which minimize the sim2real gap both in sensing and in actuation. Our agent directly predicts (discretized) velocity commands which are maintained through closed-loop control in the real robot. The behavior of the real robot (including the underlying low-level controller) is identified and simulated in a modified Habitat simulator. Noise models for odometry and localization further contribute in lowering the sim2real gap. We evaluate on real navigation scenarios explore different localization and point goal calculation methods and report significant gains in performance and robustness compared to prior work.

****

CAM Back Again: Large Kernel CNNs from a Weakly Supervised Object Localization Perspective

Shunsuke Yasuki, Masato Taki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 341-351

Recently convolutional neural networks (CNNs) with large size kernels have attracted much attention in the computer vision field following the success of the Vision Transformers. Large kernel CNNs have been reported to perform well in downstream vision tasks as well as in classification performance. The reason for the high-performance of large kernel CNNs in downstream tasks has been attributed to the large effective receptive field (ERF) produced by large size kernels but this view has not been fully tested. We therefore revisit the performance of large kernel CNNs in downstream task focusing on the weakly supervised object localization (WSOL) task. WSOL a difficult downstream task that is not fully supervised provides a new angle to explore the capabilities of the large kernel CNNs. Our study compares the modern large kernel CNNs ConvNeXt RepLKNet and SLaK to test the validity of the naive expectation that ERF size is important for improving downstream task performance. Our analysis of the factors contributing to high performance provides a different perspective in which the main factor is feature map improvement. Furthermore we find that modern CNNs are robust to the CAM problems of local regions of objects being activated which has long been discussed in WSOL. CAM is the most classic WSOL method but because of the above-mentioned problems it is often used as a baseline method for comparison. However experiments on the CUB-200-2011 dataset show that simply combining a large kernel CNN CAM and simple data augmentation methods can achieve performance (90.99% MaxBoxAcc) comparable to the latest WSOL method which is CNN-based and requires special training or complex post-processing.

****

VkD: Improving Knowledge Distillation using Orthogonal Projections

Roy Miles, Ismail Elezi, Jiankang Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15720-15730

Knowledge distillation is an effective method for training small and efficient deep learning models. However the efficacy of a single method can degenerate when transferring to other tasks modalities or even other architectures. To address this limitation we propose a novel constrained feature distillation method. This method is derived from a small set of core principles which results in two emerging components: an orthogonal projection and a task-specific normalisation. Equipped with both of these components our transformer models can outperform all previous methods on ImageNet and reach up to a 4.4% relative improvement over the previous state-of-the-art methods. To further demonstrate the generality of our

method we apply it to object detection and image generation whereby we obtain consistent and substantial performance improvements over state-of-the-art. Code and models are publicly available.
****
Putting the Object Back into Video Object Segmentation
Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, Alexander Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3151-3161
We present Cutie a video object segmentation (VOS) network with object-level memory reading which puts the object representation from memory back into the video object segmentation result. Recent works on VOS employ bottom-up pixel-level memory reading which struggles due to matching noise especially in the presence of distractors resulting in lower performance in more challenging data. In contrast Cutie performs top-down object-level memory reading by adapting a small set of object queries. Via those it interacts with the bottom-up pixel features iteratively with a query-based object transformer (qt hence Cutie). The object queries act as a high-level summary of the target object while high-resolution feature maps are retained for accurate segmentation. Together with foreground-background masked attention Cutie cleanly separates the semantics of the foreground object from the background. On the challenging MOSE dataset Cutie improves by 8.7 J&F over XMem with a similar running time and improves by 4.2 J&F over DeAOT while being three times faster. Code is available at: hkchengrex.github.io/Cutie
****
Concept Weaver: Enabling Multi-Concept Fusion in Text-to-Image Models
Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, Fabian Caba Heilbron; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8880-8889
While there has been significant progress in customizing text-to-image generation models generating images that combine multiple personalized concepts remains challenging. In this work we introduce Concept Weaver a method for composing customized text-to-image diffusion models at inference time. Specifically the method breaks the process into two steps: creating a template image aligned with the semantics of input prompts and then personalizing the template using a concept fusion strategy. The fusion strategy incorporates the appearance of the target concepts into the template image while retaining its structural details. The results indicate that our method can generate multiple custom concepts with higher identity fidelity compared to alternative approaches. Furthermore the method is shown to seamlessly handle more than two concepts and closely follow the semantic meaning of the input prompt without blending appearances across different subjects.
****
PKU-DyMVHumans: A Multi-View Video Benchmark for High-Fidelity Dynamic Human Modeling
Xiaoyun Zheng, Liwei Liao, Xufeng Li, Jianbo Jiao, Rongjie Wang, Feng Gao, Shiqi Wang, Ronggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22530-22540
High-quality human reconstruction and photo-realistic rendering of a dynamic scene is a long-standing problem in computer vision and graphics. Despite considerable efforts invested in developing various capture systems and reconstruction algorithms recent advancements still struggle with loose or oversized clothing and overly complex poses. In part this is due to the challenges of acquiring high-quality human datasets. To facilitate the development of these fields in this paper we present PKU-DyMVHumans a versatile human-centric dataset for high-fidelity reconstruction and rendering of dynamic human scenarios from dense multi-view videos. It comprises 8.2 million frames captured by more than 56 synchronized cameras across diverse scenarios. These sequences comprise 32 human subjects across 45 different scenarios each with a high-detailed appearance and realistic human motion. Inspired by recent advancements in neural radiance field (NeRF)-based scene representations we carefully set up an off-the-shelf framework that is easy to provide those state-of-the-art NeRF-based implementations and benchmark on P

KU-DyMVHumans dataset. It is paving the way for various applications like fine-g rained foreground/background decomposition high-quality human reconstruction and photo-realistic novel view synthesis of a dynamic scene. Extensive studies are performed on the benchmark demonstrating new observations and challenges that em erge from using such high-fidelity dynamic data. The project page and data is av ailable at: https://pku-dymvhumans.github.io.
****
Cross-Domain Few-Shot Segmentation via Iterative Support-Query Correspondence Mi ning
Jiahao Nie, Yun Xing, Gongjie Zhang, Pei Yan, Aoran Xiao, Yap-Peng Tan, Alex C. Kot, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2024, pp. 3380-3390
Cross-Domain Few-Shot Segmentation (CD-FSS) poses the challenge of segmenting no vel categories from a distinct domain using only limited exemplars. In this pape r we undertake a comprehensive study of CD-FSS and uncover two crucial insights: (i) the necessity of a fine-tuning stage to effectively transfer the learned me ta-knowledge across domains and (ii) the overfitting risk during the naive fine- tuning due to the scarcity of novel category examples. With these insights we pr opose a novel cross-domain fine-tuning strategy that addresses the challenging C D-FSS tasks. We first design Bi-directional Few-shot Prediction (BFP) which esta blishes support-query correspondence in a bi-directional manner crafting augment ed supervision to reduce the overfitting risk. Then we further extend BFP into I terative Few-shot Adaptor (IFA) which is a recursive framework to capture the su pport-query correspondence iteratively targeting maximal exploitation of supervi sory signals from the sparse novel category samples. Extensive empirical evaluat ions show that our method significantly outperforms the state-of-the-arts (+7.8%) which verifies that IFA tackles the cross-domain challenges and mitigates the overfitting simultaneously. The code is available at: https://github.com/niejiah ao1998/IFA.
****
CausalPC: Improving the Robustness of Point Cloud Classification by Causal Effec t Identification
Yuanmin Huang, Mi Zhang, Daizong Ding, Erling Jiang, Zhaoxiang Wang, Min Yang; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19779-19789
Deep neural networks have demonstrated remarkable performance in point cloud cla ssification. However previous works show they are vulnerable to adversarial pert urbations that can manipulate their predictions. Given the distinctive modality of point clouds various attack strategies have emerged posing challenges for exi sting defenses to achieve effective generalization. In this study we for the fir st time introduce causal modeling to enhance the robustness of point cloud class ification models. Our insight is from the observation that adversarial examples closely resemble benign point clouds from the human perspective. In our causal m odeling we incorporate two critical variables the structural information (standi ng for the key feature leading to the classification) and the hidden confounders (standing for the noise interfering with the classification). The resulting ove rall framework CausalPC consists of three sub-modules to identify the causal eff ect for robust classification. The framework is model-agnostic and adaptable for integration with various point cloud classifiers. Our approach significantly im proves the adversarial robustness of three mainstream point cloud classification models on two benchmark datasets. For instance the classification accuracy for DGCNN on ModelNet40 increases from 29.2% to 72.0% with CausalPC whereas the best -performing baseline achieves only 42.4%.
****
LASA: Instance Reconstruction from Real Scans using A Large-scale Aligned Shape Annotation Dataset
Haolin Liu, Chongjie Ye, Yinyu Nie, Yingfan He, Xiaoguang Han; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p p. 20454-20464
Instance shape reconstruction from a 3D scene involves recovering the full geome

tries of multiple objects at the semantic instance level. Many methods leverage data-driven learning due to the intricacies of scene complexity and significant indoor occlusions. Training these methods often requires a large-scale high-quality dataset with aligned and paired shape annotations with real-world scans. Existing datasets are either synthetic or misaligned restricting the performance of data-driven methods on real data. To this end we introduce LASA a Large-scale Aligned Shape Annotation Dataset comprising 10412 high-quality CAD annotations aligned with 920 real-world scene scans from ArkitScenes created manually by professional artists. On this top we propose a novel Diffusion-based Cross-Modal Shape Reconstruction (DisCo) method. It is empowered by a hybrid feature aggregation design to fuse multi-modal inputs and recover high-fidelity object geometries. Besides we present an Occupancy-Guided 3D Object Detection (OccGOD) method and demonstrate that our shape annotations provide scene occupancy clues that can further improve 3D object detection. Supported by LASA extensive experiments show that our methods achieve state-of-the-art performance in both instance-level scene reconstruction and 3D object detection tasks.
****
LaRE^2: Latent Reconstruction Error Based Method for Diffusion-Generated Image Detection
Yunpeng Luo, Junlong Du, Ke Yan, Shouhong Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17006-17015
The evolution of Diffusion Models has dramatically improved image generation quality making it increasingly difficult to differentiate between real and generated images. This development while impressive also raises significant privacy and security concerns. In response to this we propose a novel Latent REconstruction error guided feature REfinement method (LaRE^2) for detecting the diffusion-generated images. We come up with the Latent Reconstruction Error (LaRE) the first reconstruction-error based feature in the latent space for generated image detection. LaRE surpasses existing methods in terms of feature extraction efficiency while preserving crucial cues required to differentiate between the real and the fake. To exploit LaRE we propose an Error-Guided feature REfinement module (EGRE) which can refine the image feature guided by LaRE to enhance the discriminativeness of the feature. Our EGRE utilizes an align-then-refine mechanism which effectively refines the image feature for generated-image detection from both spatial and channel perspectives. Extensive experiments on the large-scale GenImage benchmark demonstrate the superiority of our LaRE^2 which surpasses the best SoTA method by up to 11.9%/12.1% average ACC/AP across 8 different image generators. LaRE also surpasses existing methods in terms of feature extraction cost delivering an impressive speed enhancement of 8 times.
****
DiffSCI: Zero-Shot Snapshot Compressive Imaging via Iterative Spectral Diffusion Model
Zhenghao Pan, Haijin Zeng, Jiezhang Cao, Kai Zhang, Yongyong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25297-25306
This paper endeavors to advance the precision of snapshot compressive imaging (SCI) reconstruction for multispectral image (MSI). To achieve this we integrate the advantageous attributes of established SCI techniques and an image generative model propose a novel structured zero-shot diffusion model dubbed DiffSCI. DiffSCI leverages the structural insights from the deep prior and optimization-based methodologies complemented by the generative capabilities offered by the contemporary denoising diffusion model. Specifically firstly we employ a pre-trained diffusion model which has been trained on a substantial corpus of RGB images as the generative denoiser within the Plug-and-Play framework for the first time. This integration allows for the successful completion of SCI reconstruction especially in the case that current methods struggle to address effectively. Secondly we systematically account for spectral band correlations and introduce a robust methodology to mitigate wavelength mismatch thus enabling seamless adaptation of the RGB diffusion model to MSIs.Thirdly an accelerated algorithm is implemented to expedite the resolution of the data subproblem. This augmentation not only a

ccelerates the convergence rate but also elevates the quality of the reconstruct ion process.We present extensive testing to show that DiffSCI exhibits discernib le performance enhancements over prevailing self-supervised and zero-shot approa ches surpassing even supervised transformer counterparts across both simulated a nd real datasets. Code is at https://github.com/PAN083/DiffSCI.
****
DiffSHEG: A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Exp ression and Gesture Generation
Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, Qifeng Chen; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 7352-7361
We propose DiffSHEG a Diffusion-based approach for Speech-driven Holistic 3D Exp ression and Gesture generation. While previous works focused on co-speech gestur e or expression generation individually the joint generation of synchronized exp ressions and gestures remains barely explored. To address this our diffusion-bas ed co-speech motion generation Transformer enables uni-directional information f low from expression to gesture facilitating improved matching of joint expressio n-gesture distributions. Furthermore we introduce an outpainting-based sampling strategy for arbitrary long sequence generation in diffusion models offering fle xibility and computational efficiency. Our method provides a practical solution that produces high-quality synchronized expression and gesture generation driven by speech. Evaluated on two public datasets our approach achieves state-of-the- art performance both quantitatively and qualitatively. Additionally a user study confirms the superiority of our method over prior approaches. By enabling the r eal-time generation of expressive and synchronized motions our method showcases its potential for various applications in the development of digital humans and embodied agents.
****
MeLFusion: Synthesizing Music from Image and Language Cues using Diffusion Model s
Sanjoy Chowdhury, Sayan Nag, K J Joseph, Balaji Vasan Srinivasan, Dinesh Manocha ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2024, pp. 26826-26835
Music is a universal language that can communicate emotions and feelings. It for ms an essential part of the whole spectrum of creative media ranging from movies to social media posts. Machine learning models that can synthesize music are pr edominantly conditioned on textual descriptions of it. Inspired by how musicians compose music not just from a movie script but also through visualizations we p ropose MeLFusion a model that can effectively use cues from a textual descriptio n and the corresponding image to synthesize music. MeLFusion is a text-to-music diffusion model with a novel "visual synapse" which effectively infuses the sema ntics from the visual modality into the generated music. To facilitate research in this area we introduce a new dataset MeLBench and propose a new evaluation me tric IMSM. Our exhaustive experimental evaluation suggests that adding visual in formation to the music synthesis pipeline significantly improves the quality of generated music measured both objectively and subjectively with a relative gain of up to 67.98% on the FAD score. We hope that our work will gather attention to this pragmatic yet relatively under-explored research area.
****
T4P: Test-Time Training of Trajectory Prediction via Masked Autoencoder and Acto r-specific Token Memory
Daehee Park, Jaeseok Jeong, Sung-Hoon Yoon, Jaewoo Jeong, Kuk-Jin Yoon; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 15065-15076
Trajectory prediction is a challenging problem that requires considering interac tions among multiple actors and the surrounding environment. While data-driven a pproaches have been used to address this complex problem they suffer from unreli able predictions under distribution shifts during test time. Accordingly several online learning methods have been proposed using regression loss from the groun d truth of observed data leveraging the auto-labeling nature of trajectory predi

ction task. We mainly tackle the following two issues. First previous works underfit and overfit as they only optimize the last layer of motion decoder. To this end we employ the masked autoencoder (MAE) for representation learning to encourage complex interaction modeling in shifted test distribution for updating deeper layers. Second utilizing the sequential nature of driving data we propose an actor-specific token memory that enables the test-time learning of actor-wise motion characteristics. Our proposed method has been validated across various challenging cross-dataset distribution shift scenarios including nuScenes Lyft Waymo and Interaction. Our method surpasses the performance of existing state-of-the-art online learning methods in terms of both prediction accuracy and computational efficiency. The code is available at https://github.com/daeheepark/T4P.
****
Noisy-Correspondence Learning for Text-to-Image Person Re-identification
Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, Peng Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27197-27206
Text-to-image person re-identification (TIReID) is a compelling topic in the cross-modal community which aims to retrieve the target person based on a textual query. Although numerous TIReID methods have been proposed and achieved promising performance they implicitly assume the training image-text pairs are correctly aligned which is not always the case in real-world scenarios. In practice the image-text pairs inevitably exist under-correlated or even false-correlated a.k.a noisy correspondence (NC) due to the low quality of the images and annotation errors. To address this problem we propose a novel Robust Dual Embedding method (RDE) that can learn robust visual-semantic associations even with NC. Specifically RDE consists of two main components: 1) A Confident Consensus Division (CCD) module that leverages the dual-grained decisions of dual embedding modules to obtain a consensus set of clean training data which enables the model to learn correct and reliable visual-semantic associations. 2) A Triplet Alignment Loss (TAL) relaxes the conventional Triplet Ranking loss with the hardest negative samples to a log-exponential upper bound over all negative ones thus preventing the model collapse under NC and can also focus on hard-negative samples for promising performance. We conduct extensive experiments on three public benchmarks namely CUHK-PEDES ICFG-PEDES and RSTPReID to evaluate the performance and robustness of our RDE. Our method achieves state-of-the-art results both with and without synthetic noisy correspondences on all three datasets. Code is available at https://github.com/QinYang79/RDE.
****
InstaGen: Enhancing Object Detection by Training on Synthetic Dataset
Chengjian Feng, Yujie Zhong, Zequn Jie, Weidi Xie, Lin Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14121-14130
In this paper we present a novel paradigm to enhance the ability of object detector e.g. expanding categories or improving detection performance by training on syn- thetic dataset generated from diffusion models. Specifically we integrate an instance-level grounding head into a pre- trained generative diffusion model to augment it with the ability of localising instances in the generated images. The grounding head is trained to align the text embedding of category names with the regional visual feature of the diffusion model using supervision from an off-the-shelf object detector and a novel self-training scheme on (novel) categories not covered by the detector. We conduct thorough experiments to show that this enhanced version of diffusion model termed as InstaGen can serve as a data synthesizer to enhance object detectors by training on its generated samples demonstrating superior performance over existing state-of-the-art methods in open-vocabulary (+4.5 AP) and data-sparse (+1.2 ? 5.2 AP) scenarios.
****
PanoRecon: Real-Time Panoptic 3D Reconstruction from Monocular Video
Dong Wu, Zike Yan, Hongbin Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21507-21518
We introduce the Panoptic 3D Reconstruction task a unified and holistic scene un

derstanding task for a monocular video. And we present PanoRecon - a novel frame work to address this new task which realizes an online geometry reconstruction a lone with dense semantic and instance labeling. Specifically PanoRecon increment ally performs panoptic 3D reconstruction for each video fragment consisting of m ultiple consecutive key frames from a volumetric feature representation using fe ed-forward neural networks. We adopt a depth-guided back-projection strategy to sparse and purify the volumetric feature representation. We further introduce a voxel clustering module to get object instances in each local fragment and then design a tracking and fusion algorithm for the integration of instances from dif ferent fragments to ensure temporal coherence. Such design enables our PanoRecon to yield a coherent and accurate panoptic 3D reconstruction. Experiments on Sca nNetV2 demonstrate a very competitive geometry reconstruction result compared wi th state-of-the-art reconstruction methods as well as promising 3D panoptic segm entation result with only RGB input while being real-time. Code is available at: https://github.com/Riser6/PanoRecon.
****
Animating General Image with Large Visual Motion Model
Dengsheng Chen, Xiaoming Wei, Xiaolin Wei; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7131-7140
We present the pioneering Large Visual Motion Model (LVMM) meticulously engineer ed to analyze the intrinsic dynamics encapsulated within real-world imagery. Our model fortified with a wealth of prior knowledge extracted from billions of ima ge pairs demonstrates promising results in predicting a diverse spectrum of scen e dynamics. As a result it can infuse any generic image with authentic dynamic e ffects enhancing its visual allure.
****
Visual Point Cloud Forecasting enables Scalable Autonomous Driving
Zetong Yang, Li Chen, Yanan Sun, Hongyang Li; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14673-14684
In contrast to extensive studies on general vision pre-training for scalable vis ual autonomous driving remains seldom explored. Visual autonomous driving applic ations require features encompassing semantics 3D geometry and temporal informat ion simultaneously for joint perception prediction and planning posing dramatic challenges for pre-training. To resolve this we bring up a new pre-training task termed as visual point cloud forecasting - predicting future point clouds from historical visual input. The key merit of this task captures the synergic learni ng of semantics 3D structures and temporal dynamics. Hence it shows superiority in various downstream tasks. To cope with this new problem we present ViDAR a ge neral model to pre-train downstream visual encoders. It first extracts historica l embeddings by the encoder. These representations are then transformed to 3D ge ometric space via a novel Latent Rendering operator for future point cloud predi ction. Experiments show significant gain in downstream tasks e.g. 3.1% NDS on 3D detection 10% error reduction on motion forecasting and 15% less collision ra te on planning.
****
Towards Transferable Targeted 3D Adversarial Attack in the Physical World
Yao Huang, Yinpeng Dong, Shouwei Ruan, Xiao Yang, Hang Su, Xingxing Wei; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24512-24522
Compared with transferable untargeted attacks transferable targeted adversarial attacks could specify the misclassification categories of adversarial samples po sing a greater threat to security-critical tasks. In the meanwhile 3D adversaria l samples due to their potential of multi-view robustness can more comprehensive ly identify weaknesses in existing deep learning systems possessing great applic ation value. However the field of transferable targeted 3D adversarial attacks r emains vacant. The goal of this work is to develop a more effective technique th at could generate transferable targeted 3D adversarial examples filling the gap in this field. To achieve this goal we design a novel framework named TT3D that could rapidly reconstruct from few multi-view images into Transferable Targeted 3D textured meshes. While existing mesh-based texture optimization methods compu

te gradients in the high-dimensional mesh space and easily fall into local optim a leading to unsatisfactory transferability and distinct distortions TT3D innovatively performs dual optimization towards both feature grid and Multi-layer Perceptron (MLP) parameters in the grid-based NeRF space which significantly enhances black-box transferability while enjoying naturalness. Experimental results show that TT3D not only exhibits superior cross-model transferability but also maintains considerable adaptability across different renders and vision tasks. More importantly we produce 3D adversarial examples with 3D printing techniques in the real world and verify their robust performance under various scenarios.
****
SwitchLight: Co-design of Physics-driven Architecture and Pre-training Framework for Human Portrait Relighting
Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, Sanghyun Woo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25096-25106
We introduce a co-designed approach for human portrait relighting that combines a physics-guided architecture with a pre-training framework. Drawing on the Cook-Torrance reflectance model we have meticulously configured the architecture design to precisely simulate light-surface interactions. Furthermore to overcome the limitation of scarce high-quality lightstage data we have developed a self-supervised pre-training strategy. This novel combination of accurate physical modeling and expanded training dataset establishes a new benchmark in relighting realism.
****
DIRECT-3D: Learning Direct Text-to-3D Generation on Massive Noisy 3D Data
Qihao Liu, Yi Zhang, Song Bai, Adam Kortylewski, Alan Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6881-6891
We present DIRECT-3D a diffusion-based 3D generative model for creating high-quality 3D assets (represented by Neural Radiance Fields) from text prompts. Unlike recent 3D generative models that rely on clean and well-aligned 3D data limiting them to single or few-class generation our model is directly trained on extensive noisy and unaligned `in-the-wild' 3D assets mitigating the key challenge (i.e. data scarcity) in large-scale 3D generation. In particular DIRECT-3D is a tri-plane diffusion model that integrates two innovations: 1) A novel learning framework where noisy data are filtered and aligned automatically during the training process. Specifically after an initial warm-up phase using a small set of clean data an iterative optimization is introduced in the diffusion process to explicitly estimate the 3D pose of objects and select beneficial data based on conditional density. 2) An efficient 3D representation that is achieved by disentangling object geometry and color features with two separate conditional diffusion models that are optimized hierarchically. Given a prompt input our model generates high-quality high-resolution realistic and complex 3D objects with accurate geometric details in seconds. We achieve state-of-the-art performance in both single-class generation and text-to-3D generation. We also demonstrate that DIRECT-3D can serve as a useful 3D geometric prior of objects for example to alleviate the well-known Janus problem in 2D-lifting methods such as DreamFusion.
****
Synthesize Step-by-Step: Tools Templates and LLMs as Data Generators for Reasoning-Based Chart VQA
Zhuowan Li, Bhavan Jasani, Peng Tang, Shabnam Ghadar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13613-13623
Understanding data visualizations like charts and plots requires reasoning about both visual elements and numerics. Although strong in extractive questions current chart visual question answering (chart VQA) models suffer on complex reasoning questions. In this work we address the lack of reasoning ability by data augmentation. We leverage Large Language Models (LLMs) which have shown to have strong reasoning ability as an automatic data annotator that generates question-answer annotations for chart images. The key innovation in our method lies in the Sy

nthesize Step-by-Step strategy: our LLM-based data generator learns to decompose the complex question into step-by-step sub-questions (rationales) which are then used to derive the final answer using external tools i.e. Python. This step-wise generation procedure is trained on synthetic data generated using a template-based QA generation pipeline. Experimental results highlight the significance of the proposed step-by-step generation. By training with the LLM-augmented data (LAMENDA) we significantly enhance the chart VQA models achieving the state-of-the-art accuracy on the ChartQA and PlotQA datasets. In particular our approach improves the accuracy of the previous state-of-the-art approach from 38% to 54% on the human-written questions in the ChartQA dataset which needs strong reasoning. We hope our work underscores the potential of synthetic data and encourages further exploration of data augmentation using LLMs for reasoning-heavy tasks.
****
LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding
Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, Cong Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15630-15640
Recently leveraging large language models (LLMs) or multimodal large language models (MLLMs) for document understanding has been proven very promising. However previous works that employ LLMs/MLLMs for document understanding have not fully explored and utilized the document layout information which is vital for precise document understanding. In this paper we propose LayoutLLM an LLM/MLLM based method for document understanding. The core of LayoutLLM is a layout instruction tuning strategy which is specially designed to enhance the comprehension and utilization of document layouts. The proposed layout instruction tuning strategy consists of two components: Layout-aware Pre-training and Layout-aware Supervised Fine-tuning. To capture the characteristics of document layout in Layout-aware Pre-training three groups of pre-training tasks corresponding to document-level region-level and segment-level information are introduced. Furthermore a novel module called layout chain-of-thought (LayoutCoT) is devised to enable LayoutLLM to focus on regions relevant to the question and generate accurate answers. LayoutCoT is effective for boosting the performance of document understanding. Meanwhile it brings a certain degree of interpretability which could facilitate manual inspection and correction. Experiments on standard benchmarks show that the proposed LayoutLLM significantly outperforms existing methods that adopt open-source 7B LLMs/MLLMs for document understanding.
****
ProTeCt: Prompt Tuning for Taxonomic Open Set Classification
Tz-Ying Wu, Chih-Hui Ho, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16531-16540
Visual-language foundation models like CLIP learn generalized representations that enable zero-shot open-set classification. Few-shot adaptation methods based on prompt tuning have been shown to further improve performance on downstream datasets. However these methods do not fare well in the taxonomic open set (TOS) setting where the classifier is asked to make prediction from label set across different levels of semantic granularity. Frequently they infer incorrect labels at coarser taxonomic class levels even when the inference at the leaf level (original class labels) is correct. To address this problem we propose a prompt tuning technique that calibrates the hierarchical consistency of model predictions. A set of metrics of hierarchical consistency the Hierarchical Consistent Accuracy (HCA) and the Mean Treecut Accuracy (MTA) are first proposed to evaluate TOS model performance. A new Prompt Tuning for Hierarchical Consistency (ProTeCt) technique is then proposed to calibrate classification across label set granularities. Results show that ProTeCt can be combined with existing prompt tuning methods to significantly improve TOS classification without degrading the leaf level classification performance.
****
Adapters Strike Back
Jan-Martin O. Steitz, Stefan Roth; Proceedings of the IEEE/CVF Conference on Com

puter Vision and Pattern Recognition (CVPR), 2024, pp. 23449-23459
Adapters provide an efficient and lightweight mechanism for adapting trained transformer models to a variety of different tasks. However they have often been found to be outperformed by other adaptation mechanisms including low-rank adaptation. In this paper we provide an in-depth study of adapters their internal structure as well as various implementation choices. We uncover pitfalls for using adapters and suggest a concrete improved adapter architecture called Adapter+ that not only outperforms previous adapter implementations but surpasses a number of other more complex adaptation mechanisms in several challenging settings. Despite this our suggested adapter is highly robust and unlike previous work requires little to no manual intervention when addressing a novel scenario. Adapter+ reaches state-of-the-art average accuracy on the VTAB benchmark even without a per-task hyperparameter optimization.
****

Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology
Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Dominique Beaini, Maciej Sypetkowski, Chi Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, Berton Earnshaw; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11757-11768
Featurizing microscopy images for use in biological research remains a significant challenge especially for large-scale experiments spanning millions of images. This work explores the scaling properties of weakly supervised classifiers and self-supervised masked autoencoders (MAEs) when training with increasingly larger model backbones and microscopy datasets. Our results show that ViT-based MAEs outperform weakly supervised classifiers on a variety of tasks achieving as much as a 11.5% relative improvement when recalling known biological relationships curated from public databases. Additionally we develop a new channel-agnostic MAE architecture (CA-MAE) that allows for inputting images of different numbers and orders of channels at inference time. We demonstrate that CA-MAEs effectively generalize by inferring and evaluating on a microscopy image dataset (JUMP-CP) generated under different experimental conditions with a different channel structure than our pretraining data (RPI-93M). Our findings motivate continued research into scaling self-supervised learning on microscopy data in order to create powerful foundation models of cellular biology that have the potential to catalyze advancements in drug discovery and beyond.
****

OHTA: One-shot Hand Avatar via Data-driven Implicit Priors
Xiaozheng Zheng, Chao Wen, Zhuo Su, Zeran Xu, Zhaohu Li, Yang Zhao, Zhou Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 799-810
In this paper we delve into the creation of one-shot hand avatars attaining high-fidelity and drivable hand representations swiftly from a single image. With the burgeoning domains of the digital human the need for quick and personalized hand avatar creation has become increasingly critical. Existing techniques typically require extensive input data and may prove cumbersome or even impractical in certain scenarios. To enhance accessibility we present a novel method OHTA (One-shot Hand avaTAr) that enables the creation of detailed hand avatars from merely one image. OHTA tackles the inherent difficulties of this data-limited problem by learning and utilizing data-driven hand priors. Specifically we design a hand prior model initially employed for 1) learning various hand priors with available data and subsequently for 2) the inversion and fitting of the target identity with prior knowledge. OHTA demonstrates the capability to create high-fidelity hand avatars with consistent animatable quality solely relying on a single image. Furthermore we illustrate the versatility of OHTA through diverse applications encompassing text-to-avatar conversion hand editing and identity latent space manipulation.
****

Segment and Caption Anything
Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan

Wang, Zicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13405-13417
We propose a method to efficiently equip the Segment Anything Model (SAM) with the ability to generate regional captions. SAM presents strong generalizability to segment anything while is short for semantic understanding. By introducing a lightweight query-based feature mixer we align the region-specific features with the embedding space of language models for later caption generation. As the number of trainable parameters is small (typically in the order of tens of millions) it costs less computation less memory usage and less communication bandwidth resulting in both fast and scalable training. To address the scarcity problem of regional caption data we propose to first pre-train our model on objection detection and segmentation tasks. We call this step weak supervision pretraining since the pretraining data only contains category names instead of full-sentence descriptions. The weak supervision pretraining allows us to leverage many publicly available object detection and segmentation datasets. We conduct extensive experiments to demonstrate the superiority of our method and validate each design choice. This work serves as a stepping stone towards scaling up regional captioning data and sheds light on exploring efficient ways to augment SAM with regional semantics.
****
Human Motion Prediction Under Unexpected Perturbation
Jiangbei Yue, Baiyi Li, Julien Pettré, Armin Seyfried, He Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1501-1511
We investigate a new task in human motion prediction which is predicting motions under unexpected physical perturbation potentially involving multiple people. Compared with existing research this task involves predicting less controlled unpremeditated and pure reactive motions in response to external impact and how such motions can propagate through people. It brings new challenges such as data scarcity and predicting complex interactions. To this end we propose a new method capitalizing differentiable physics and deep neural networks leading to an explicit Latent Differentiable Physics (LDP) model. Through experiments we demonstrate that LDP has high data efficiency outstanding prediction accuracy strong generalizability and good explainability. Since there is no similar research a comprehensive comparison with 11 adapted baselines from several relevant domains is conducted showing LDP outperforming existing research both quantitatively and qualitatively improving prediction accuracy by as much as 70% and demonstrating significantly stronger generalization.
****
Text-to-3D Generation with Bidirectional Diffusion using both 2D and 3D priors
Lihe Ding, Shaocong Dong, Zhanpeng Huang, Zibin Wang, Yiyuan Zhang, Kaixiong Gong, Dan Xu, Tianfan Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5115-5124
Most 3D generation research focuses on up-projecting 2D foundation models into the 3D space either by minimizing 2D Score Distillation Sampling (SDS) loss or fine-tuning on multi-view datasets. Without explicit 3D priors these methods often lead to geometric anomalies and multi-view inconsistency. Recently researchers have attempted to improve the genuineness of 3D objects by directly training on 3D datasets albeit at the cost of low-quality texture generation due to the limited texture diversity in 3D datasets. To harness the advantages of both approaches we propose Bidirectional Diffusion (BiDiff) a unified framework that incorporates both a 3D and a 2D diffusion process to preserve both 3D fidelity and 2D texture richness respectively. Moreover as a simple combination may yield inconsistent generation results we further bridge them with novel bidirectional guidance. In addition our method can be used as an initialization of optimization-based models to further improve the quality of 3D model and efficiency of optimization reducing the process from 3.4 hours to 20 minutes. Experimental results have shown that our model achieves high-quality diverse and scalable 3D generation. Project website https://bidiff.github.io/.
****

CLIP-Driven Open-Vocabulary 3D Scene Graph Generation via Cross-Modality Contras
tive Learning
Lianggangxu Chen, Xuejiao Wang, Jiale Lu, Shaohui Lin, Changbo Wang, Gaoqi He; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 27863-27873
3D Scene Graph Generation (3DSGG) aims to classify objects and their predicates
within 3D point cloud scenes. However current 3DSGG methods struggle with two ma
in challenges. 1) The dependency on labor-intensive ground-truth annotations. 2)
 Closed-set classes training hampers the recognition of novel objects and predic
ates. Addressing these issues our idea is to extract cross-modality features by
CLIP from text and image data naturally related to 3D point clouds. Cross-modali
ty features are used to train a robust 3D scene graph (3DSG) feature extractor.
Specifically we propose a novel Cross-Modality Contrastive Learning 3DSGG (CCL-3
DSGG) method. Firstly to align the text with 3DSG the text is parsed into word l
evel that are consistent with the 3DSG annotation. To enhance robustness during
the alignment adjectives are exchanged for different objects as negative samples
. Then to align the image with 3DSG the camera view is treated as a positive sam
ple and other views as negatives. Lastly the recognition of novel object and pre
dicate classes is achieved by calculating the cosine similarity between prompts
and 3DSG features. Our rigorous experiments confirm the superior open-vocabulary
 capability and applicability of CCL-3DSGG in real-world contexts.
****
Adversarial Backdoor Attack by Naturalistic Data Poisoning on Trajectory Predict
ion in Autonomous Driving
Mozhgan Pourkeshavarz, Mohammad Sabokrou, Amir Rasouli; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1488
5-14894
In autonomous driving behavior prediction is fundamental for safe motion plannin
g hence the security and robustness of prediction models against adversarial att
acks are of paramount importance. We propose a novel adversarial backdoor attack
 against trajectory prediction models as a means of studying their potential vul
nerabilities. Our attack affects the victim at training time via naturalistic he
nce stealthy poisoned samples crafted using a novel two-step approach. First the
 triggers are crafted by perturbing the trajectory of attacking vehicle and then
 disguised by transforming the scene using a bi-level optimization technique. Th
e proposed attack does not depend on a particular model architecture and operate
s in a black-box manner thus can be effective without any knowledge of the victi
m model. We conduct extensive empirical studies using state-of-the-art predictio
n models on two benchmark datasets using metrics customized for trajectory predi
ction. We show that the proposed attack is highly effective as it can significan
tly hinder the performance of prediction models unnoticeable by the victims and
efficient as it forces the victim to generate malicious behavior even under cons
trained conditions. Via ablative studies we analyze the impact of different atta
ck design choices followed by an evaluation of existing defence mechanisms again
st the proposed attack.
****
Make-It-Vivid: Dressing Your Animatable Biped Cartoon Characters from Text
Junshu Tang, Yanhong Zeng, Ke Fan, Xuheng Wang, Bo Dai, Kai Chen, Lizhuang Ma; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 6243-6253
Creating and animating 3D biped cartoon characters is crucial and valuable in va
rious applications. Compared with geometry the diverse texture design plays an i
mportant role in making 3D biped cartoon characters vivid and charming. Therefor
e we focus on automatic texture design for cartoon characters based on input ins
tructions. This is challenging for domain-specific requirements and a lack of hi
gh-quality data. To address this challenge we propose Make-It-Vivid the first at
tempt to enable high-quality texture generation from text in UV space. We prepar
e a detailed text-texture paired data for 3D characters by using vision-question
-answering agents. Then we customize a pretrained text-to-image model to generat
e texture map with template structure while preserving the natural 2D image know

ledge. Furthermore to enhance fine-grained details we propose a novel adversaria
l learning scheme to shorten the domain gap between original dataset and realist
ic texture domain. Extensive experiments show that our approach outperforms curr
ent texture generation methods resulting in efficient character texturing and fa
ithful generation with prompts. Besides we showcase various applications such as
 out of domain generation and texture stylization. We also provide an efficient
generation system for automatic text-guided textured character generation and an
imation.
****
StraightPCF: Straight Point Cloud Filtering
Dasith de Silva Edirimuni, Xuequan Lu, Gang Li, Lei Wei, Antonio Robles-Kelly, H
ongdong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2024, pp. 20721-20730
Point cloud filtering is a fundamental 3D vision task which aims to remove noise
 while recovering the underlying clean surfaces. State-of-the-art methods remove
 noise by moving noisy points along stochastic trajectories to the clean surface
s. These methods often require regularization within the training objective and/
or during post-processing to ensure fidelity. In this paper we introduce Straigh
tPCF a new deep learning based method for point cloud filtering. It works by mov
ing noisy points along straight paths thus reducing discretization errors while
ensuring faster convergence to the clean surfaces. We model noisy patches as int
ermediate states between high noise patch variants and their clean counterparts
and design the VelocityModule to infer a constant flow velocity from the former
to the latter. This constant flow leads to straight filtering trajectories. In a
ddition we introduce a DistanceModule that scales the straight trajectory using
an estimated distance scalar to attain convergence near the clean surface. Our n
etwork is lightweight and only has 530K parameters being 17% of IterativePFN (a
 most recent point cloud filtering network). Extensive experiments on both synth
etic and real-world data show our method achieves state-of-the-art results. Our
method also demonstrates nice distributions of filtered points without the need
for regularization. The implementation code can be found at: https://github.com/
ddsediri/StraightPCF.
****
Mirasol3B: A Multimodal Autoregressive Model for Time-Aligned and Contextual Mod
alities
AJ Piergiovanni, Isaac Noble, Dahun Kim, Michael S. Ryoo, Victor Gomes, Anelia A
ngelova; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2024, pp. 26804-26814
One of the main challenges of multimodal learning is the need to combine heterog
eneous modalities (e.g. video audio text). For example video and audio are obtai
ned at much higher rates than text and are roughly aligned in time. They are oft
en not synchronized with text which comes as a global context e.g. a title or a
description. Furthermore video and audio inputs are of much larger volumes and g
row as the video length increases which naturally requires more compute dedicate
d to these modalities and makes modeling of long-range dependencies harder. We h
ere decouple the multimodal modeling dividing it into separate autoregressive mo
dels processing the inputs according to the characteristics of the modalities. W
e propose a multimodal model consisting of an autoregressive component for the t
ime-synchronized modalities (audio and video) and an autoregressive component fo
r the context modalities which are not necessarily aligned in time but are still
 sequential. To address the long-sequences of the video-audio inputs we further
partition the video and audio sequences in consecutive snippets and autoregressi
vely process their representations. To that end we propose a Combiner mechanism
which models the audio-video information jointly producing compact but expressiv
e representations. This allows us to scale to 512 input video frames without inc
rease in model parameters. Our approach achieves the state-of-the-art on multipl
e well established multimodal benchmarks. It effectively addresses the high comp
utational demand of media inputs by learning compact representations controlling
 the sequence length of the audio-video feature representations and modeling the
ir dependencies in time.

****
Neural Sign Actors: A Diffusion Model for 3D Sign Language Production from Text
Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1985-1995
Sign Languages (SL) serve as the primary mode of communication for the Deaf and Hard of Hearing communities. Deep learning methods for SL recognition and translation have achieved promising results. However Sign Language Production (SLP) poses a challenge as the generated motions must be realistic and have precise semantic meaning. Most SLP methods rely on 2D data which hinders their realism. In this work a diffusion-based SLP model is trained on a curated large-scale dataset of 4D signing avatars and their corresponding text transcripts. The proposed method can generate dynamic sequences of 3D avatars from an unconstrained domain of discourse using a diffusion process formed on a novel and anatomically informed graph neural network defined on the SMPL-X body skeleton. Through quantitative and qualitative experiments we show that the proposed method considerably outperforms previous methods of SLP. This work makes an important step towards realistic neural sign avatars bridging the communication gap between Deaf and hearing communities.
****
On the Diversity and Realism of Distilled Dataset: An Efficient Dataset Distillation Paradigm
Peng Sun, Bei Shi, Daiwei Yu, Tao Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9390-9399
Contemporary machine learning which involves training large neural networks on massive datasets faces significant computational challenges. Dataset distillation as a recent emerging strategy aims to compress real-world datasets for efficient training. However this line of research currently struggles with large-scale and high-resolution datasets hindering its practicality and feasibility. Thus we re-examine existing methods and identify three properties essential for real-world applications: realism diversity and efficiency. As a remedy we propose RDED a novel computationally-efficient yet effective data distillation paradigm to enable both diversity and realism of the distilled data. Extensive empirical results over various model architectures and datasets demonstrate the advancement of RDED: we can distill a dataset to 10 images per class from full ImageNet-1K within 7 minutes achieving a notable 42% accuracy with ResNet-18 on a single RTX-4090 GPU (while the SOTA only achieves 21% but requires 6 hours). Code: https://github.com/LINs-lab/RDED.
****
Semantics-aware Motion Retargeting with Vision-Language Models
Haodong Zhang, Zhike Chen, Haocheng Xu, Lei Hao, Xiaofei Wu, Songcen Xu, Zhensong Zhang, Yue Wang, Rong Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2155-2164
Capturing and preserving motion semantics is essential to motion retargeting between animation characters. However most of the previous works neglect the semantic information or rely on human-designed joint-level representations. Here we present a novel Semantics-aware Motion reTargeting (SMT) method with the advantage of vision-language models to extract and maintain meaningful motion semantics. We utilize a differentiable module to render 3D motions. Then the high-level motion semantics are incorporated into the motion retargeting process by feeding the vision-language model with the rendered images and aligning the extracted semantic embeddings. To ensure the preservation of fine-grained motion details and high-level semantics we adopt a two-stage pipeline consisting of skeleton-aware pre-training and fine-tuning with semantics and geometry constraints. Experimental results show the effectiveness of the proposed method in producing high-quality motion retargeting results while accurately preserving motion semantics. Project page can be found at https://sites.google.com/view/smtnet.
****
Semantically-Shifted Incremental Adapter-Tuning is A Continual ViTransformer
Yuwen Tan, Qinhao Zhou, Xiang Xiang, Ke Wang, Yuchuan Wu, Yongbin Li; Proceeding

s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23252-23262
Class-incremental learning (CIL) aims to enable models to continuously learn new classes while overcoming catastrophic forgetting. The introduction of pre-trained models has brought new tuning paradigms to CIL. In this paper we revisit different parameter-efficient tuning (PET) methods within the context of continual learning. We observe that adapter tuning demonstrates superiority over prompt-based methods even without parameter expansion in each learning session. Motivated by this we propose incrementally tuning the shared adapter without imposing parameter update constraints enhancing the learning capacity of the backbone. Additionally we employ feature sampling from stored prototypes to retrain a unified classifier further improving its performance. We estimate the semantic shift of old prototypes without access to past samples and update stored prototypes session by session. Our proposed method eliminates model expansion and avoids retaining any image samples. It surpasses previous pre-trained model-based CIL methods and demonstrates remarkable continual learning capabilities. Experimental results on five CIL benchmarks validate the effectiveness of our approach achieving state-of-the-art (SOTA) performance.
****
Low-Rank Approximation for Sparse Attention in Multi-Modal LLMs
Lin Song, Yukang Chen, Shuai Yang, Xiaohan Ding, Yixiao Ge, Ying-Cong Chen, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13763-13773
This paper focuses on the high computational complexity in Large Language Models (LLMs) a significant challenge in both natural language processing (NLP) and multi-modal tasks. We propose Low-Rank Approximation for Sparse At- tention (LoRA-Sparse) an innovative approach that strate- gically reduces this complexity. LoRA-Sparse introduces low-rank linear projection layers for sparse attention ap- proximation. It utilizes an order-mimic training methodol- ogy which is crucial for efficiently approximating the self- attention mechanism in LLMs. We empirically show that sparse attention not only reduces computational demands but also enhances model performance in both NLP and multi-modal tasks. This surprisingly shows that redundant attention in LLMs might be non-beneficial. We extensively validate LoRA-Sparse through rigorous empirical studies in both (NLP) and multi-modal tasks demonstrating its effec- tiveness and general applicability. Based on LLaMA and LLaVA models our methods can reduce more than half of the self-attention computation with even better performance than full-attention baselines.
****
TASeg: Temporal Aggregation Network for LiDAR Semantic Segmentation
Xiaopei Wu, Yuenan Hou, Xiaoshui Huang, Binbin Lin, Tong He, Xinge Zhu, Yuexin Ma, Boxi Wu, Haifeng Liu, Deng Cai, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15311-15320
Training deep models for LiDAR semantic segmentation is challenging due to the inherent sparsity of point clouds. Utilizing temporal data is a natural remedy against the sparsity problem as it makes the input signal denser. However previous multi-frame fusion algorithms fall short in utilizing sufficient temporal information due to the memory constraint and they also ignore the informative temporal images. To fully exploit rich information hidden in long-term temporal point clouds and images we present the Temporal Aggregation Network termed TASeg. Specifically we propose a Temporal LiDAR Aggregation and Distillation (TLAD) algorithm which leverages historical priors to assign different aggregation steps for different classes. It can largely reduce memory and time overhead while achieving higher accuracy. Besides TLAD trains a teacher injected with gt priors to distill the model further boosting the performance. To make full use of temporal images we design a Temporal Image Aggregation and Fusion (TIAF) module which can greatly expand the camera FOV and enhance the present features. Temporal LiDAR points in the camera FOV are used as mediums to transform temporal image features to the present coordinate for temporal multi-modal fusion. Moreover we develop a Static-Moving Switch Augmentation (SMSA) algorithm which utilizes sufficient temporal information to enable objects to switch their motion states freely thus grea

tly increasing static and moving training samples. Our TASeg ranks 1st on three challenging tracks i.e. SemanticKITTI single-scan track multi-scan track and nuScenes LiDAR segmentation track strongly demonstrating the superiority of our method. Codes are available at https://github.com/LittlePey/TASeg.
****
Bootstrapping SparseFormers from Vision Foundation Models
Ziteng Gao, Zhan Tong, Kevin Qinghong Lin, Joya Chen, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17710-17721
The recently proposed SparseFormer architecture provides an alternative approach to visual understanding by utilizing a significantly lower number of visual tokens via adjusting RoIs greatly reducing computational costs while still achieving promising performance. However training SparseFormers from scratch is still expensive and scaling up the number of parameters can be challenging. In this paper we propose to bootstrap SparseFormers from ViT-based vision foundation models in a simple and efficient way. Since the majority of SparseFormer blocks are the standard transformer ones we can inherit weights from large-scale pre-trained vision transformers and freeze them as much as possible. Therefore we only need to train the SparseFormer-specific lightweight focusing transformer to adjust token RoIs and fine-tune a few early pre-trained blocks to align the final token representation. In such a way we can bootstrap SparseFormer architectures from various large-scale pre-trained models (e.g. IN-21K pre-trained AugRegs or CLIPs) using a rather smaller amount of training samples (e.g. IN-1K) and without labels or captions within just a few hours. As a result the bootstrapped unimodal SparseFormer (from AugReg-ViT-L/16-384) can reach 84.9% accuracy on IN-1K with only 49 tokens and the multimodal SparseFormer from CLIPs also demonstrates notable zero-shot performance with highly reduced computational cost without seeing any caption during the bootstrapping procedure. In addition CLIP-bootstrapped SparseFormers which align the output space with language without seeing a word can serve as efficient vision encoders in multimodal large language models. Code and models are available at https://github.com/showlab/sparseformer
****
EventPS: Real-Time Photometric Stereo Using an Event Camera
Bohan Yu, Jieji Ren, Jin Han, Feishi Wang, Jinxiu Liang, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9602-9611
Photometric stereo is a well-established technique to estimate the surface normal of an object. However the requirement of capturing multiple high dynamic range images under different illumination conditions limits the speed and real-time applications. This paper introduces EventPS a novel approach to real-time photometric stereo using an event camera. Capitalizing on the exceptional temporal resolution dynamic range and low bandwidth characteristics of event cameras EventPS estimates surface normal only from the radiance changes significantly enhancing data efficiency. EventPS seamlessly integrates with both optimization-based and deep-learning-based photometric stereo techniques to offer a robust solution for non-Lambertian surfaces. Extensive experiments validate the effectiveness and efficiency of EventPS compared to frame-based counterparts. Our algorithm runs at over 30 fps in real-world scenarios unleashing the potential of EventPS in time-sensitive and high-speed downstream applications.
****
Unsupervised Semantic Segmentation Through Depth-Guided Feature Correlation and Sampling
Leon Sick, Dominik Engel, Pedro Hermosilla, Timo Ropinski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3637-3646
Traditionally training neural networks to perform semantic segmentation requires expensive human-made annotations. But more recently advances in the field of unsupervised learning have made significant progress on this issue and towards closing the gap to supervised algorithms. To achieve this semantic knowledge is distilled by learning to correlate randomly sampled features from images across an

entire dataset. In this work we build upon these advances by incorporating infor
mation about the structure of the scene into the training process through the us
e of depth information. We achieve this by (1) learning depth-feature correlatio
n by spatially correlating the feature maps with the depth maps to induce knowle
dge about the structure of the scene and (2) exploiting farthest-point sampling
to more effectively select relevant features by utilizing 3D sampling techniques
 on depth information of the scene. Finally we demonstrate the effectiveness of
our technical contributions through extensive experimentation and present signif
icant improvements in performance across multiple benchmark datasets.
****
On the Road to Portability: Compressing End-to-End Motion Planner for Autonomous
 Driving
Kaituo Feng, Changsheng Li, Dongchun Ren, Ye Yuan, Guoren Wang; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 15099-15108
End-to-end motion planning models equipped with deep neural networks have shown
great potential for enabling full autonomous driving. However the oversized neur
al networks render them impractical for deployment on resource-constrained syste
ms which unavoidably requires more computational time and resources during refer
ence. To handle this knowledge distillation offers a promising approach that com
presses models by enabling a smaller student model to learn from a larger teache
r model. Nevertheless how to apply knowledge distillation to compress motion pla
nners has not been explored so far. In this paper we propose PlanKD the first kn
owledge distillation framework tailored for compressing end-to-end motion planne
rs. First considering that driving scenes are inherently complex often containin
g planning-irrelevant or even noisy information transferring such information is
 not beneficial for the student planner. Thus we design an information bottlenec
k based strategy to only distill planning-relevant information rather than trans
fer all information indiscriminately. Second different waypoints in an output pl
anned trajectory may hold varying degrees of importance for motion planning wher
e a slight deviation in certain crucial waypoints might lead to a collision. The
refore we devise a safety-aware waypoint-attentive distillation module that assi
gns adaptive weights to different waypoints based on the importance to encourage
 the student to accurately mimic more crucial waypoints thereby improving overal
l safety. Experiments demonstrate that our PlanKD can boost the performance of s
maller planners by a large margin and significantly reduce their reference time.
****
RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diff
usion Models
Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M. Rehg, Pinar Yanardag; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 6507-6516
Recent advancements in diffusion-based models have demonstrated significant succ
ess in generating images from text. However video editing models have not yet re
ached the same level of visual quality and user control. To address this we intr
oduce RAVE a zero-shot video editing method that leverages pre-trained text-to-i
mage diffusion models without additional training. RAVE takes an input video and
 a text prompt to produce high-quality videos while preserving the original moti
on and semantic structure. It employs a novel noise shuffling strategy leveragin
g spatio-temporal interactions between frames to produce temporally consistent v
ideos faster than existing methods. It is also efficient in terms of memory requ
irements allowing it to handle longer videos. RAVE is capable of a wide range of
 edits from local attribute modifications to shape transformations. In order to
demonstrate the versatility of RAVE we create a comprehensive video evaluation d
ataset ranging from object-focused scenes to complex human activities like danci
ng and typing and dynamic scenes featuring swimming fish and boats. Our qualitat
ive and quantitative experiments highlight the effectiveness of RAVE in diverse
video editing scenarios compared to existing methods. Our code dataset and video
s can be found in \href https://rave-video-edit.github.io/.
****

PredToken: Predicting Unknown Tokens and Beyond with Coarse-to-Fine Iterative De
coding
Xuesong Nie, Haoyuan Jin, Yunfeng Yan, Xi Chen, Zhihang Zhu, Donglian Qi; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2024, pp. 18143-18152
Predictive learning models which aim to predict future frames based on past obse
rvations are crucial to constructing world models. These models need to maintain
 low-level consistency and capture high-level dynamics in unannotated spatiotemp
oral data. Transitioning from frame-wise to token-wise prediction presents a via
ble strategy for addressing these needs. How to improve token representation and
 optimize token decoding presents significant challenges. This paper introduces
PredToken a novel predictive framework that addresses these issues by decoupling
 space-time tokens into distinct components for iterative cascaded decoding. Con
cretely we first design a "decomposition quantization and reconstruction" schema
 based on VQGAN to improve the token representation. This scheme disentangles lo
w- and high-frequency representations and employs a dimension-aware quantization
 model allowing more low-level details to be preserved. Building on this we pres
ent a "coarse-to-fine iterative decoding" method. It leverages dynamic soft deco
ding to refine coarse tokens and static soft decoding for fine tokens enabling m
ore high-level dynamics to be captured. These designs make PredToken produce hig
h-quality predictions. Extensive experiments demonstrate the superiority of our
method on various real-world spatiotemporal predictive benchmarks. Furthermore P
redToken can also be extended to other visual generative tasks to yield realisti
c outcomes.
****
Video-Based Human Pose Regression via Decoupled Space-Time Aggregation
Jijie He, Wenwu Yang; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2024, pp. 1022-1031
By leveraging temporal dependency in video sequences multi-frame human pose esti
mation algorithms have demonstrated remarkable results in complicated situations
 such as occlusion motion blur and video defocus. These algorithms are predomina
ntly based on heatmaps resulting in high computation and storage requirements pe
r frame which limits their flexibility and real-time application in video scenar
ios particularly on edge devices. In this paper we develop an efficient and effe
ctive video-based human pose regression method which bypasses intermediate repre
sentations such as heatmaps and instead directly maps the input to the output jo
int coordinates. Despite the inherent spatial correlation among adjacent joints
of the human pose the temporal trajectory of each individual joint exhibits rela
tive independence. In light of this we propose a novel Decoupled Space-Time Aggr
egation network (DSTA) to separately capture the spatial contexts between adjace
nt joints and the temporal cues of each individual joint thereby avoiding the co
nflation of spatiotemporal dimensions. Concretely DSTA learns a dedicated featur
e token for each joint to facilitate the modeling of their spatiotemporal depend
encies. With the proposed joint-wise local-awareness attention mechanism our met
hod is capable of efficiently and flexibly utilizing the spatial dependency of a
djacent joints and the temporal dependency of each joint itself. Extensive exper
iments demonstrate the superiority of our method. Compared to previous regressio
n-based single-frame human pose estimation methods DSTA significantly enhances p
erformance achieving an 8.9 mAP improvement on PoseTrack2017. Furthermore our ap
proach either surpasses or is on par with the state-of-the-art heatmap-based mul
ti-frame human pose estimation methods. Project page: https://github.com/zgspose
/DSTA.
****
L-MAGIC: Language Model Assisted Generation of Images with Coherence
Zhipeng Cai, Matthias Mueller, Reiner Birkl, Diana Wofk, Shao-Yen Tseng, Junda C
heng, Gabriela Ben-Melech Stan, Vasudev Lai, Michael Paulitsch; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 7049-7058
In the current era of generative AI breakthroughs generating panoramic scenes fr
om a single input image remains a key challenge. Most existing methods use diffu

sion-based iterative or simultaneous multi-view inpainting. However the lack of global scene layout priors leads to subpar outputs with duplicated objects (e.g. multiple beds in a bedroom) or requires time-consuming human text inputs for each view. We propose L-MAGIC a novel method leveraging large language models for guidance while diffusing multiple coherent views of 360 degree panoramic scenes. L-MAGIC harnesses pre-trained diffusion and language models without fine-tuning ensuring zero-shot performance. The output quality is further enhanced by super-resolution and multi-view fusion techniques. Extensive experiments demonstrate that the resulting panoramic scenes feature better scene layouts and perspective view rendering quality compared to related works with >70% preference in human evaluations. Combined with conditional diffusion models L-MAGIC can accept various input modalities including but not limited to text depth maps sketches and colored scripts. Applying depth estimation further enables 3D point cloud generation and dynamic scene exploration with fluid camera motion.
****
3D Face Tracking from 2D Video through Iterative Dense UV to Image Flow
Felix Taubner, Prashant Raina, Mathieu Tuli, Eu Wern Teh, Chul Lee, Jinmiao Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1227-1237
When working with 3D facial data improving fidelity and avoiding the uncanny valley effect is critically dependent on accurate 3D facial performance capture. Because such methods are expensive and due to the widespread availability of 2D videos recent methods have focused on how to perform monocular 3D face tracking. However these methods often fall short in capturing precise facial movements due to limitations in their network architecture training and evaluation processes. Addressing these challenges we propose a novel face tracker FlowFace that introduces an innovative 2D alignment network for dense per-vertex alignment. Unlike prior work FlowFace is trained on high-quality 3D scan annotations rather than weak supervision or synthetic data. Our 3D model fitting module jointly fits a 3D face model from one or many observations integrating existing neutral shape priors for enhanced identity and expression disentanglement and per-vertex deformations for detailed facial feature reconstruction. Additionally we propose a novel metric and benchmark for assessing tracking accuracy. Our method exhibits superior performance on both custom and publicly available benchmarks. We further validate the effectiveness of our tracker by generating high-quality 3D data from 2D videos which leads to performance gains on downstream tasks.
****
Carve3D: Improving Multi-view Reconstruction Consistency for Diffusion Models with RL Finetuning
Desai Xie, Jiahao Li, Hao Tan, Xin Sun, Zhixin Shu, Yi Zhou, Sai Bi, Sören Pirk, Arie E. Kaufman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6369-6379
Multi-view diffusion models obtained by applying Supervised Finetuning (SFT) to text-to-image diffusion models have driven recent breakthroughs in text-to-3D research. However due to the limited size and quality of existing 3D datasets they still suffer from multi-view inconsistencies and Neural Radiance Field (NeRF) reconstruction artifacts. We argue that multi-view diffusion models can benefit from further Reinforcement Learning Finetuning (RLFT) which allows models to learn from the data generated by themselves and improve beyond their dataset limitations during SFT. To this end we introduce Carve3D an improved RLFT algorithm coupled with a novel Multi-view Reconstruction Consistency (MRC) metric to enhance the consistency of multi-view diffusion models. To measure the MRC metric on a set of multi-view images we compare them with their corresponding NeRF renderings at the same camera viewpoints. The resulting model which we denote as Carve3DM demonstrates superior multi-view consistency and NeRF reconstruction quality than existing models. Our results suggest that pairing SFT with Carve3D's RLFT is essential for developing multi-view-consistent diffusion models mirroring the standard Large Language Model (LLM) alignment pipeline. Our code training and testing data and video results are available at: https://desaixie.github.io/carve-3d.
****

Random Entangled Tokens for Adversarially Robust Vision Transformer
Huihui Gong, Minjing Dong, Siqi Ma, Seyit Camtepe, Surya Nepal, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24554-24563
Vision Transformers (ViTs) have emerged as a compelling alternative to Convolutional Neural Networks (CNNs) in the realm of computer vision showcasing tremendous potential. However recent research has unveiled a susceptibility of ViTs to adversarial attacks akin to their CNN counterparts. Adversarial training and randomization are two representative effective defenses for CNNs. Some researchers have attempted to apply adversarial training to ViTs and achieved comparable robustness to CNNs while it is not easy to directly apply randomization to ViTs because of the architecture difference between CNNs and ViTs. In this paper we delve into the structural intricacies of ViTs and propose a novel defense mechanism termed Random entangled image Transformer (ReiT) which seamlessly integrates adversarial training and randomization to bolster the adversarial robustness of ViTs. Recognizing the challenge posed by the structural disparities between ViTs and CNNs we introduce a novel module input-independent random entangled self-attention (II-ReSA). This module optimizes random entangled tokens that lead to "dissimilar" self-attention outputs by leveraging model parameters and the sampled random tokens thereby synthesizing the self-attention module outputs and random entangled tokens to diminish adversarial similarity. ReiT incorporates two distinct random entangled tokens and employs dual randomization offering an effective countermeasure against adversarial examples while ensuring comprehensive deduction guarantees. Through extensive experiments conducted on various ViT variants and benchmarks we substantiate the superiority of our proposed method in enhancing the adversarial robustness of Vision Transformers.
****
Shadow Generation for Composite Image Using Diffusion Model
Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, Li Niu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8121-8130
In the realm of image composition generating realistic shadow for the inserted foreground remains a formidable challenge. Previous works have developed image-to-image translation models which are trained on paired training data. However they are struggling to generate shadows with accurate shapes and intensities hindered by data scarcity and inherent task complexity. In this paper we resort to foundation model with rich prior knowledge of natural shadow images. Specifically we first adapt ControlNet to our task and then propose intensity modulation modules to improve the shadow intensity. Moreover we extend the small-scale DESOBA dataset to DESOBAv2 using a novel data acquisition pipeline. Experimental results on both DESOBA and DESOBAv2 datasets as well as real composite images demonstrate the superior capability of our model for shadow generation task. The dataset code and model are released at https://github.com/bcmi/Object-Shadow-Generation-Dataset-DESOBAv2.
****
DisCo: Disentangled Control for Realistic Human Dance Generation
Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, Lijuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9326-9336
Generative AI has made significant strides in computer vision particularly in text-driven image/video synthesis (T2I/T2V). Despite the notable advancements it remains challenging in human-centric content synthesis such as realistic dance generation. Current methodologies primarily tailored for human motion transfer encounter difficulties when confronted with real-world dance scenarios (e.g. social media dance) which require to generalize across a wide spectrum of poses and intricate human details. In this paper we depart from the traditional paradigm of human motion transfer and emphasize two additional critical attributes for the synthesis of human dance content in social media contexts: (i) Generalizability: the model should be able to generalize beyond generic human viewpoints as well as unseen human subjects backgrounds and poses; (ii) Compositionality: it should

allow for the seamless composition of seen/unseen subjects backgrounds and poses from different sources. To address these challenges we introduce DISCO which includes a novel model architecture with disentangled control to improve the compositionality of dance synthesis and an effective human attribute pre-training for better generalizability to unseen humans. Extensive qualitative and quantitative results demonstrate that DISCO can generate high-quality human dance images and videos with diverse appearances and flexible motions. Code is available at https://disco-dance.github.io/.
****
L2B: Learning to Bootstrap Robust Models for Combating Label Noise
Yuyin Zhou, Xianhang Li, Fengze Liu, Qingyue Wei, Xuxi Chen, Lequan Yu, Cihang Xie, Matthew P. Lungren, Lei Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23523-23533
Deep neural networks have shown great success in representation learning. Deep neural networks have shown great success in representation learning. However when learning with noisy labels (LNL) they can easily overfit and fail to generalize to new data. This paper introduces a simple and effective method named Learning to Bootstrap (L2B) which enables models to bootstrap themselves using their own predictions without being adversely affected by erroneous pseudo-labels. It achieves this by dynamically adjusting the importance weight between real observed and generated labels as well as between different samples through meta-learning. Unlike existing instance reweighting methods the key to our method lies in a new versatile objective that enables implicit relabeling concurrently leading to significant improvements without incurring additional costs. L2B offers several benefits over the baseline methods. It yields more robust models that are less susceptible to the impact of noisy labels by guiding the bootstrapping procedure more effectively. It better exploits the valuable information contained in corrupted instances by adapting the weights of both instances and labels. Furthermore L2B is compatible with existing LNL methods and delivers competitive results spanning natural and medical imaging tasks including classification and segmentation under both synthetic and real-world noise. Extensive experiments demonstrate that our method effectively mitigates the challenges of noisy labels often necessitating few to no validation samples and is well generalized to other tasks such as image segmentation. This not only positions it as a robust complement to existing LNL techniques but also underscores its practical applicability. The code and models are available at https://github.com/yuyinzhou/l2b.
****
GaussianShader: 3D Gaussian Splatting with Shading Functions for Reflective Surfaces
Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, Yuexin Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5322-5332
The advent of neural 3D Gaussians has recently brought about a revolution in the field of neural rendering facilitating the generation of high-quality renderings at real-time speeds. However the explicit and discrete representation encounters challenges when applied to scenes featuring reflective surfaces. In this paper we present GaussianShader a novel method that applies a simplified shading function on 3D Gaussians to enhance the neural rendering in scenes with reflective surfaces while preserving the training and rendering efficiency. The main challenge in applying the shading function lies in the accurate normal estimation on discrete 3D Gaussians. Specifically we proposed a novel normal estimation framework based on the shortest axis directions of 3D Gaussians with a delicately designed loss to make the consistency between the normals and the geometries of Gaussian spheres. Experiments show that GaussianShader strikes a commendable balance between efficiency and visual quality. Our method surpasses Gaussian Splatting in PSNR on specular object datasets exhibiting an improvement of 1.57dB. When compared to prior works handling reflective surfaces such as Ref-NeRF our optimization time is significantly accelerated (23h vs. 0.58h).
****
Tactile-Augmented Radiance Fields

Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, Andrew Owens; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26529-26539
We present a scene representation that brings vision and touch into a shared 3D space which we call a tactile-augmented radiance field. This representation capitalizes on two key insights: (i) ubiquitous vision-based touch sensors are built on perspective cameras and (ii) visually and structurally similar regions of a scene share the same tactile features. We use these insights to train a conditional diffusion model that provided with an RGB image and a depth map rendered from a neural radiance field generates its corresponding tactile "image". To train this diffusion model we collect the largest collection of spatially-aligned visual and tactile data. Through qualitative and quantitative experiments we demonstrate the accuracy of our cross-modal generative model and the utility of collected and rendered visual-tactile pairs across a range of downstream tasks. Project page: https://dou-yiming.github.io/TaRF
****
Intensity-Robust Autofocus for Spike Camera
Changqing Su, Zhiyuan Ye, Yongsheng Xiao, You Zhou, Zhen Cheng, Bo Xiong, Zhaofei Yu, Tiejun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25018-25027
Spike cameras a novel neuromorphic visual sensor can capture full-time spatial information through spike stream offering ultra-high temporal resolution and an extensive dynamic range. Autofocus control (AC) plays a pivotal role in a camera to efficiently capture information in challenging real-world scenarios. Nevertheless due to disparities in data modality and information characteristics compared to frame stream and event stream the current lack of efficient AC methods has made it challenging for spike cameras to adapt to intricate real-world conditions. To address this challenge we introduce a spike-based autofocus framework that includes a spike-specific focus measure called spike dispersion (SD) which effectively mitigates the influence of variations in scene light intensity during the focusing process by leveraging the spike camera's ability to record full-time spatial light intensity. Additionally the framework integrates a fast search strategy called spike-based golden fast search (SGFS) allowing rapid focal positioning without the need for a complete focus range traversal. To validate the performance of our method we have collected a spike-based autofocus dataset (SAD) containing synthetic data and real-world data under varying scene brightness and motion scenarios. Experimental results on these datasets demonstrate that our method offers state-of-the-art accuracy and efficiency. Furthermore experiments with data captured under varying scene brightness levels illustrate the robustness of our method to changes in light intensity during the focusing process.
****
FairCLIP: Harnessing Fairness in Vision-Language Learning
Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, Yi Fang, Mengyu Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12289-12301
Fairness is a critical concern in deep learning especially in healthcare where these models influence diagnoses and treatment decisions. Although fairness has been investigated in the vision-only domain the fairness of medical vision-language (VL) models remains unexplored due to the scarcity of medical VL datasets for studying fairness. To bridge this research gap we introduce the first fair vision-language medical dataset (Harvard-FairVLMed) that provides detailed demographic attributes ground-truth labels and clinical notes to facilitate an in-depth examination of fairness within VL foundation models. Using Harvard-FairVLMed we conduct a comprehensive fairness analysis of two widely-used VL models (CLIP and BLIP2) pre-trained on both natural and medical domains across four different protected attributes. Our results highlight significant biases in all VL models with Asian Male Non-Hispanic and Spanish being the preferred subgroups across the protected attributes of race gender ethnicity and language respectively. In order to alleviate these biases we propose FairCLIP an optimal-transport-based approa

ch that achieves a favorable trade-off between performance and fairness by reducing the Sinkhorn distance between the overall sample distribution and the distributions corresponding to each demographic group. As the first VL dataset of its kind Harvard-FairVLMed holds the potential to catalyze advancements in the development of machine learning models that are both ethically aware and clinically effective. Our dataset and code are available at https://ophai.hms.harvard.edu/datasets/harvard-fairvlmed10k.
****
StreamingFlow: Streaming Occupancy Forecasting with Asynchronous Multi-modal Data Streams via Neural Ordinary Differential Equation
Yining Shi, Kun Jiang, Ke Wang, Jiusi Li, Yunlong Wang, Mengmeng Yang, Diange Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14833-14842
Predicting the future occupancy states of the surrounding environment is a vital task for autonomous driving. However current best-performing single-modality methods or multi-modality fusion perception methods are only able to predict uniform snapshots of future occupancy states and require strictly synchronized sensory data for sensor fusion. We propose a novel framework StreamingFlow to lift these strong limitations. StreamingFlow is a novel BEV occupancy predictor that ingests asynchronous multi-sensor data streams for fusion and performs streaming forecasting of the future occupancy map at any future timestamps. By integrating neural ordinary differential equations (N-ODE) into recurrent neural networks StreamingFlow learns derivatives of BEV features over temporal horizons updates the implicit sensor's BEV features as part of the fusion process and propagates BEV states to the desired future time point. It shows good zero-shot generalization ability of prediction reflected in the interpolation of the observed prediction time horizon and the reasonable inference of the unseen farther future period. Extensive experiments on two large-scale datasets nuScenes and Lyft L5 demonstrate that StreamingFlow significantly outperforms previous vision-based LiDAR-based methods and shows superior performance compared to state-of-the-art fusion-based methods.
****
pix2gestalt: Amodal Segmentation by Synthesizing Wholes
Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3931-3940
We introduce pix2gestalt a framework for zero-shot amodal segmentation which learns to estimate the shape and appearance of whole objects that are only partially visible behind occlusions. By capitalizing on large-scale diffusion models and transferring their representations to this task we learn a conditional diffusion model for reconstructing whole objects in challenging zero-shot cases including examples that break natural and physical priors such as art. As training data we use a synthetically curated dataset containing occluded objects paired with their whole counterparts. Experiments show that our approach outperforms supervised baselines on established benchmarks. Our model can furthermore be used to significantly improve the performance of existing object recognition and 3D reconstruction methods in the presence of occlusions.
****
Weakly Supervised Point Cloud Semantic Segmentation via Artificial Oracle
Hyeokjun Kweon, Jihun Kim, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3721-3731
Manual annotation of every point in a point cloud is a costly and labor-intensive process. While weakly supervised point cloud semantic segmentation (WSPCSS) with sparse annotation shows promise the limited information from initial sparse labels can place an upper bound on performance. As a new research direction for WSPCSS we propose a novel Region Exploration via Artificial Labeling (REAL) framework. It leverages a foundational image model as an artificial oracle within the active learning context eliminating the need for manual annotation by a human oracle. To integrate the 2D model into the 3D domain we first introduce a Projection-based Point-toSegment (PP2S) module designed to enable prompt segmentation o

f 3D data without additional training. The REAL framework samples query points based on model predictions and requests annotations from PP2S dynamically refining labels and improving model training. Furthermore to overcome several challenges of employing an artificial model as an oracle we formulate effective query sampling and label updating strategies. Our comprehensive experiments and comparisons demonstrate that the REAL framework significantly outperforms existing methods across various benchmarks. The code is available at https://github.com/jihun1998/AO.
****

Language Model Guided Interpretable Video Action Reasoning
Ning Wang, Guangming Zhu, HS Li, Liang Zhang, Syed Afaq Ali Shah, Mohammed Bennamoun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18878-18887
Although neural networks excel in video action recognition tasks their "black-box" nature makes it challenging to understand the rationale behind their decisions. Recent approaches used inherently interpretable models to analyze video actions in a manner akin to human reasoning. However it has been observed that these interpretable models tend to underperform when compared to their black-box counterparts. In this work we present a new framework called Language-guided Interpretable Action Recognition framework (LaIAR). This framework leverages knowledge from language models to enhance both the recognition capabilities and the interpretability of video models. In essence we reframe the challenge of understanding video model decisions as a task of aligning video and language models. Using the logical reasoning captured by the language model we steer the training of the video model. This integrated approach not only improves the video model's adaptability to different domains but also boosts its overall performance. Extensive experiments on Charades and CAD-120 datasets demonstrate the superior performance and interpretability of our proposed method. The code of LaIAR is available at https://github.com/NingWang2049/LaIAR.
****

Forecasting of 3D Whole-body Human Poses with Grasping Objects
Haitao Yan, Qiongjie Cui, Jiexin Xie, Shijie Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1726-1736
In the context of computer vision and human-robot interaction forecasting 3D human poses is crucial for understanding human behavior and enhancing the predictive capabilities of intelligent systems. While existing methods have made significant progress they often focus on predicting major body joints overlooking fine-grained gestures and their interaction with objects. Human hand movements particularly during object interactions play a pivotal role and provide more precise expressions of human poses. This work fills this gap and introduces a novel paradigm: forecasting 3D whole-body human poses with a focus on grasping objects. This task involves predicting activities across all joints in the body and hands encompassing the complexities of internal heterogeneity and external interactivity. To tackle these challenges we also propose a novel approach: C^3HOST cross-context cross-modal consolidation for 3D whole-body pose forecasting effectively handles the complexities of internal heterogeneity and external interactivity. C^3HOST involves distinct steps including the heterogeneous content encoding and alignment and cross-modal feature learning and interaction. These enable us to predict activities across all body and hand joints ensuring high-precision whole-body human pose prediction even during object grasping. Extensive experiments on two benchmarks demonstrate that our model significantly enhances the accuracy of whole-body human motion prediction. The project page is available at https://sites.google.com/view/c3host.
****

COTR: Compact Occupancy TRansformer for Vision-based 3D Occupancy Prediction
Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, Yuan Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19936-19945
The autonomous driving community has shown significant interest in 3D occupancy prediction driven by its exceptional geometric perception and general object rec

ognition capabilities. To achieve this current works try to construct a Tri-Pers pective View (TPV) or Occupancy (OCC) representation extending from the Bird-Eye -View perception. However compressed views like TPV representation lose 3D geome try information while raw and sparse OCC representation requires heavy but redun dant computational costs. To address the above limitations we propose Compact Oc cupancy TRansformer (COTR) with a geometry-aware occupancy encoder and a semanti c-aware group decoder to reconstruct a compact 3D OCC representation. The occupa ncy encoder first generates a compact geometrical OCC feature through efficient explicit-implicit view transformation. Then the occupancy decoder further enhanc es the semantic discriminability of the compact OCC representation by a coarse-t o-fine semantic grouping strategy. Empirical experiments show that there are evi dent performance gains across multiple baselines e.g. COTR outperforms baselines with a relative improvement of 8%-15% demonstrating the superiority of our meth od.
****
Accelerating Diffusion Sampling with Optimized Time Steps
Shuchen Xue, Zhaoqiang Liu, Fei Chen, Shifeng Zhang, Tianyang Hu, Enze Xie, Zhen guo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2024, pp. 8292-8301
Diffusion probabilistic models (DPMs) have shown remarkable performance in high-resolution image synthesis but their sampling efficiency is still to be desired due to the typically large number of sampling steps. Recent advancements in high -order numerical ODE solvers for DPMs have enabled the generation of high-qualit y images with much fewer sampling steps. While this is a significant development most sampling methods still employ uniform time steps which is not optimal when using a small number of steps. To address this issue we propose a general frame work for designing an optimization problem that seeks more appropriate time step s for a specific numerical ODE solver for DPMs. This optimization problem aims t o minimize the distance between the ground-truth solution to the ODE and an appr oximate solution corresponding to the numerical solver. It can be efficiently so lved using the constrained trust region method taking less than 15 seconds. Our extensive experiments on both unconditional and conditional sampling using pixel - and latent-space DPMs demonstrate that when combined with the state-of-the-art sampling method UniPC our optimized time steps significantly improve image gene ration performance in terms of FID scores for datasets such as CIFAR-10 and Imag eNet compared to using uniform time steps.
****
See Say and Segment: Teaching LMMs to Overcome False Premises
Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang , Joseph E. Gonzalez, Trevor Darrell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13459-13469
Current open-source Large Multimodal Models (LMMs) excel at tasks such as open-v ocabulary language grounding and segmentation but can suffer under false premise s when queries imply the existence of something that is not actually present in the image. We observe that existing methods that fine-tune an LMM to segment ima ges significantly degrade their ability to reliably determine ("see") if an obje ct is present and to interact naturally with humans ("say") a form of catastroph ic forgetting. In this work we propose a cascading and joint training approach f or LMMs to solve this task avoiding catastrophic forgetting of previous skills. Our resulting model can "see" by detecting whether objects are present in an ima ge "say" by telling the user if they are not proposing alternative queries or co rrecting semantic errors in the query and finally "segment" by outputting the ma sk of the desired objects if they exist. Additionally we introduce a novel False Premise Correction benchmark dataset an extension of existing RefCOCO(+/g) refe rring segmentation datasets (which we call FP-RefCOCO(+/g)). The results show th at our method not only detects false premises up to 55% better than existing app roaches but under false premise conditions produces relative cIOU improvements o f more than 31% over baselines and produces natural language feedback judged hel pful up to 67% of the time.
****

Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving?
Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, Jose M. Alvarez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14864-14873
End-to-end autonomous driving recently emerged as a promising research direction to target autonomy from a full-stack perspective. Along this line many of the latest works follow an open-loop evaluation setting on nuScenes to study the planning behavior. In this paper we delve deeper into the problem by conducting thorough analyses and demystifying more devils in the details. We initially observed that the nuScenes dataset characterized by relatively simple driving scenarios leads to an under-utilization of perception information in end-to-end models incorporating ego status such as the ego vehicle's velocity. These models tend to rely predominantly on the ego vehicle's status for future path planning. Beyond the limitations of the dataset we also note that current metrics do not comprehensively assess the planning quality leading to potentially biased conclusions drawn from existing benchmarks. To address this issue we introduce a new metric to evaluate whether the predicted trajectories adhere to the road. We further propose a simple baseline able to achieve competitive results without relying on perception annotations. Given the current limitations on the benchmark and metrics we suggest the community reassess relevant prevailing research and be cautious about whether the continued pursuit of state-of-the-art would yield convincing and universal conclusions. Code and models are available at https://github.com/NVlabs/BEV-Planner.
****
Unsupervised Template-assisted Point Cloud Shape Correspondence Network
Jiacheng Deng, Jiahao Lu, Tianzhu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5250-5259
Unsupervised point cloud shape correspondence aims to establish point-wise correspondences between source and target point clouds. Existing methods obtain correspondences directly by computing point-wise feature similarity between point clouds. However non-rigid objects possess strong deformability and unusual shapes making it a longstanding challenge to directly establish correspondences between point clouds with unconventional shapes. To address this challenge we propose an unsupervised Template-Assisted point cloud shape correspondence Network termed TANet including a template generation module and a template assistance module. The proposed TANet enjoys several merits. Firstly the template generation module establishes a set of learnable templates with explicit structures. Secondly we introduce a template assistance module that extensively leverages the generated templates to establish more accurate shape correspondences from multiple perspectives. Extensive experiments on four human and animal datasets demonstrate that TANet achieves favorable performance against state-of-the-art methods.
****
CGI-DM: Digital Copyright Authentication for Diffusion Models via Contrasting Gradient Inversion
Xiaoyu Wu, Yang Hua, Chumeng Liang, Jiaru Zhang, Hao Wang, Tao Song, Haibing Guan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10812-10821
Diffusion Models (DMs) have evolved into advanced image generation tools especially for few-shot generation where a pre-trained model is fine-tuned on a small set of images to capture a specific style or object. Despite their success concerns exist about potential copyright violations stemming from the use of unauthorized data in this process. In response we present Contrasting Gradient Inversion for Diffusion Models (CGI-DM) a novel method featuring vivid visual representations for digital copyright authentication. Our approach involves removing partial information of an image and recovering missing details by exploiting conceptual differences between the pre-trained and fine-tuned models. We formulate the differences as KL divergence between latent variables of the two models when given the same input image which can be maximized through Monte Carlo sampling and Projected Gradient Descent (PGD). The similarity between original and recovered images serves as a strong indicator of potential infringements. Extensive experimen

ts on the WikiArt and Dreambooth datasets demonstrate the high accuracy of CGI-DM in digital copyright authentication surpassing alternative validation techniques. Code implementation is available at https://github.com/Nicholas0228/Revelio.
****

Making Visual Sense of Oracle Bones for You and Me
Runqi Qiao, Lan Yang, Kaiyue Pang, Honggang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12656-12665
Visual perception evolves over time. This is particularly the case of oracle bone scripts where visual glyphs seem intuitive to people from distant past prove difficult to be understood in contemporary eyes. While semantic correspondence of an oracle can be found via a dictionary lookup this proves to be not enough for public viewers to connect the dots i.e. why does this oracle mean that? Common solution relies on a laborious curation process to collect visual guide for each oracle (Fig.1) which hinges on the case-by-case effort and taste of curators. This paper delves into one natural follow-up question: can AI take over?Begin with a comprehensive human study we show participants could indeed make better sense of an oracle glyph subjected to a proper visual guide and its efficacy can be approximated via a novel metric termed TransOV (Transferable Oracle Visuals). We then define a new conditional visual generation task based on an oracle glyph and its semantic meaning and importantly approach it by circumventing any form of model training in the presence of fatal lack of oracle data. At its heart is to leverage foundation model like GPT-4V to reason about the visual cues hidden inside an oracle and take advantage of an existing text-to-image model for final visual guide generation. Extensive empirical evidence shows our AI-enabled visual guides achieve significantly comparable TransOV performance compared with those collected under manual efforts. Finally we demonstrate the versatility of our system under a more complex setting where it is required to work alongside an AI image denoiser to cope with raw oracle scan image inputs (cf. processed clean oracle glyphs). Code is available at https://github.com/RQ-Lab/OBS-Visual.
****

Finsler-Laplace-Beltrami Operators with Application to Shape Analysis
Simon Weber, Thomas Dagès, Maolin Gao, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3131-3140
The Laplace-Beltrami operator (LBO) emerges from studying manifolds equipped with a Riemannian metric. It is often called the swiss army knife of geometry processing as it allows to capture intrinsic shape information and gives rise to heat diffusion geodesic distances and a multitude of shape descriptors. It also plays a central role in geometric deep learning. In this work we explore Finsler manifolds as a generalization of Riemannian manifolds. We revisit the Finsler heat equation and derive a Finsler heat kernel and a Finsler-Laplace-Beltrami Operator (FLBO): a novel theoretically justified anisotropic Laplace-Beltrami operator (ALBO). In experimental evaluations we demonstrate that the proposed FLBO is a valuable alternative to the traditional Riemannian-based LBO and ALBOs for spatial filtering and shape correspondence estimation. We hope that the proposed Finsler heat kernel and the FLBO will inspire further exploration of Finsler geometry in the computer vision community.
****

Minimal Perspective Autocalibration
Andrea Porfiri Dal Cin, Timothy Duff, Luca Magri, Tomas Pajdla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5064-5073
We introduce a new family of minimal problems for reconstruction from multiple views. Our primary focus is a novel approach to autocalibration a long-standing problem in computer vision. Traditional approaches to this problem such as those based on Kruppa's equations or the modulus constraint rely explicitly on the knowledge of multiple fundamental matrices or a projective reconstruction. In contrast we consider a novel formulation involving constraints on image points the unknown depths of 3D points and a partially specified calibration matrix K. For 2

and 3 views we present a comprehensive taxonomy of minimal autocalibration probl
ems obtained by relaxing some of these constraints. These problems are organized
 into classes according to the number of views and any assumed prior knowledge o
f K. Within each class we determine problems with the fewest---or a relatively s
mall number of---solutions. From this zoo of problems we devise three practical
solvers. Experiments with synthetic and real data and interfacing our solvers wi
th COLMAP demonstrate that we achieve superior accuracy compared to state-of-the
-art calibration methods. The code is available at https://github.com/andreadalc
in/MinimalPerspectiveAutocalibration.
****
MOHO: Learning Single-view Hand-held Object Reconstruction with Multi-view Occlu
sion-Aware Supervision
Chenyangguang Zhang, Guanlong Jiao, Yan Di, Gu Wang, Ziqin Huang, Ruida Zhang, F
abian Manhardt, Bowen Fu, Federico Tombari, Xiangyang Ji; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 99
92-10002
Previous works concerning single-view hand-held object reconstruction typically
rely on supervision from 3D ground-truth models which are hard to collect in rea
l world. In contrast readily accessible hand-object videos offer a promising tra
ining data source but they only give heavily occluded object observations. In th
is paper we present a novel synthetic-to-real framework to exploit Multi-view Oc
clusion-aware supervision from hand-object videos for Hand-held Object reconstru
ction (MOHO) from a single image tackling two predominant challenges in such set
ting: hand-induced occlusion and object's self-occlusion. First in the synthetic
 pre-training stage we render a large-scaled synthetic dataset SOMVideo with han
d-object images and multi-view occlusion-free supervisions adopted to address ha
nd-induced occlusion in both 2D and 3D spaces. Second in the real-world finetuni
ng stage MOHO leverages the amodal-mask-weighted geometric supervision to mitiga
te the unfaithful guidance caused by the hand-occluded supervising views in real
 world. Moreover domain-consistent occlusion-aware features are amalgamated in M
OHO to resist object's self-occlusion for inferring the complete object shape. E
xtensive experiments on HO3D and DexYCB datasets demonstrate 2D-supervised MOHO
gains superior results against 3D-supervised methods by a large margin.
****
BANF: Band-Limited Neural Fields for Levels of Detail Reconstruction
Akhmedkhan Shabanov, Shrisudhan Govindarajan, Cody Reading, Lily Goli, Daniel Re
bain, Kwang Moo Yi, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20571-20580
Largely due to their implicit nature neural fields lack a direct mechanism for f
iltering as Fourier analysis from discrete signal processing is not directly app
licable to these representations. Effective filtering of neural fields is critic
al to enable level-of-detail processing in downstream applications and support o
perations that involve sampling the field on regular grids (e.g. marching cubes)
. Existing methods that attempt to decompose neural fields in the frequency doma
in either resort to heuristics or require extensive modifications to the neural
field architecture. We show that via a simple modification one can obtain neural
 fields that are low-pass filtered and in turn show how this can be exploited to
 obtain a frequency decomposition of the entire signal. We demonstrate the valid
ity of our technique by investigating level-of-detail reconstruction and showing
 how coarser representations can be computed effectively.
****
Time- Memory- and Parameter-Efficient Visual Adaptation
Otniel-Bogdan Mercea, Alexey Gritsenko, Cordelia Schmid, Anurag Arnab; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2024, pp. 5536-5545
As foundation models become more popular there is a growing need to efficiently
finetune them for downstream tasks. Although numerous adaptation methods have be
en proposed they are designed to be efficient only in terms of how many paramete
rs are trained. They however typically still require backpropagating gradients t
hroughout the model meaning that their training-time and -memory cost does not r

educe as significantly. We propose an adaptation method which does not backpropagate gradients through the backbone. We achieve this by designing a lightweight network in parallel that operates on features from the frozen pretrained backbone. As a result our method is efficient not only in terms of parameters but also in training-time and memory usage. Our approach achieves state-of-the-art accuracy-parameter trade-offs on the popular VTAB benchmark and we further show how we outperform prior works with respect to training-time and -memory usage too. We further demonstrate the training efficiency and scalability of our method by adapting a vision transformer backbone of 4 billion parameters for the computationally demanding task of video classification without any intricate model parallelism. Here we outperform a prior adaptor-based method which could only scale to a 1 billion parameter backbone or fully-finetuning a smaller backbone with the same GPU and less training time.
****
SecondPose: SE(3)-Consistent Dual-Stream Feature Fusion for Category-Level Pose Estimation
Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, Benjamin Busam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9959-9969
Category-level object pose estimation aiming to predict the 6D pose and 3D size of objects from known categories typically struggles with large intra-class shape variation. Existing works utilizing mean shapes often fall short of capturing this variation. To address this issue we present SecondPose a novel approach integrating object-specific geometric features with semantic category priors from DINOv2. Leveraging the advantage of DINOv2 in providing SE(3)-consistent semantic features we hierarchically extract two types of SE(3)-invariant geometric features to further encapsulate local-to-global object-specific information. These geometric features are then point-aligned with DINOv2 features to establish a consistent object representation under SE(3) transformations facilitating the mapping from camera space to the pre-defined canonical space thus further enhancing pose estimation. Extensive experiments on NOCS-REAL275 demonstrate that SecondPose achieves a 12.4% leap forward over the state-of-the-art. Moreover on a more complex dataset HouseCat6D which provides photometrically challenging objects SecondPose still surpasses other competitors by a large margin. Code is released at https://github.com/NOrangeeroli/SecondPose.git.
****
Physical Property Understanding from Language-Embedded Feature Fields
Albert J. Zhai, Yuan Shen, Emily Y. Chen, Gloria X. Wang, Xinlei Wang, Sheng Wang, Kaiyu Guan, Shenlong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28296-28305
Can computers perceive the physical properties of objects solely through vision? Research in cognitive science and vision science has shown that humans excel at identifying materials and estimating their physical properties based purely on visual appearance. In this paper we present a novel approach for dense prediction of the physical properties of objects using a collection of images. Inspired by how humans reason about physics through vision we leverage large language models to propose candidate materials for each object. We then construct a language-embedded point cloud and estimate the physical properties of each 3D point using a zero-shot kernel regression approach. Our method is accurate annotation-free and applicable to any object in the open world. Experiments demonstrate the effectiveness of the proposed approach in various physical property reasoning tasks such as estimating the mass of common objects as well as other properties like friction and hardness.
****
EgoGen: An Egocentric Synthetic Data Generator
Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, Siyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14497-14509
Understanding the world in first-person view is fundamental in Augmented Reality

(AR). This immersive perspective brings dramatic visual changes and unique chal
lenges compared to third-person views. Synthetic data has empowered third-person
-view vision models but its application to embodied egocentric perception tasks
remains largely unexplored. A critical challenge lies in simulating natural huma
n movements and behaviors that effectively steer the embodied cameras to capture
 a faithful egocentric representation of the 3D world. To address this challenge
 we introduce EgoGen a new synthetic data generator that can produce accurate an
d rich ground-truth training data for egocentric perception tasks. At the heart
of EgoGen is a novel human motion synthesis model that directly leverages egocen
tric visual inputs of a virtual human to sense the 3D environment. Combined with
 collision-avoiding motion primitives and a two-stage reinforcement learning app
roach our motion synthesis model offers a closed-loop solution where the embodie
d perception and movement of the virtual human are seamlessly coupled. Compared
to previous works our model eliminates the need for a pre-defined global path an
d is directly applicable to dynamic environments. Combined with our easy-to-use
and scalable data generation pipeline we demonstrate EgoGen's efficacy in three
tasks: mapping and localization for head-mounted cameras egocentric camera track
ing and human mesh recovery from egocentric views. EgoGen will be fully open-sou
rced offering a practical solution for creating realistic egocentric training da
ta and aiming to serve as a useful tool for egocentric computer vision research.
****
Suppress and Rebalance: Towards Generalized Multi-Modal Face Anti-Spoofing
Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Wenzhong Tang, Zitong Yu,
 Alex Kot; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2024, pp. 211-221
Face Anti-Spoofing (FAS) is crucial for securing face recognition systems agains
t presentation attacks. With advancements in sensor manufacture and multi-modal
learning techniques many multi-modal FAS approaches have emerged. However they f
ace challenges in generalizing to unseen attacks and deployment conditions. Thes
e challenges arise from (1) modality unreliability where some modality sensors l
ike depth and infrared undergo significant domain shifts in varying environments
 leading to the spread of unreliable information during cross-modal feature fusi
on and (2) modality imbalance where training overly relies on a dominant modalit
y hinders the convergence of others reducing effectiveness against attack types
that are indistinguishable by sorely using the dominant modality. To address mod
ality unreliability we propose the Uncertainty-Guided Cross-Adapter (U-Adapter)
to recognize unreliably detected regions within each modality and suppress the i
mpact of unreliable regions on other modalities. For modality imbalance we propo
se a Rebalanced Modality Gradient Modulation (ReGrad) strategy to rebalance the
convergence speed of all modalities by adaptively adjusting their gradients. Bes
ides we provide the first large-scale benchmark for evaluating multi-modal FAS p
erformance under domain generalization scenarios. Extensive experiments demonstr
ate that our method outperforms state-of-the-art methods. Source codes and proto
cols are released on https://github.com/OMGGGGG/mmdg.
****
LEAD: Exploring Logit Space Evolution for Model Selection
Zixuan Hu, Xiaotong Li, Shixiang Tang, Jun Liu, Yichun Hu, Ling-Yu Duan; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 28664-28673
The remarkable success of "pretrain-then-finetune" paradigm has led to a prolife
ration of available pre-trained models for vision tasks. This surge presents a s
ignificant challenge in efficiently choosing the most suitable pre-trained model
s for downstream tasks. The critical aspect of this challenge lies in effectivel
y predicting the model transferability by considering the underlying fine-tuning
 dynamics. Existing methods often model fine-tuning dynamics in feature space wi
th linear transformations which do not precisely align with the fine-tuning obje
ctive and fail to grasp the essential nonlinearity from optimization. To this en
d we present LEAD a finetuning-aligned approach based on the network output of l
ogits. LEAD proposes a theoretical framework to model the optimization process a
nd derives an ordinary differential equation (ODE) to depict the nonlinear evolu

tion toward the final logit state. Additionally we design a class-aware decompos
ition method to consider the varying evolution dynamics across classes and furth
er ensure practical applicability. Integrating the closely aligned optimization
objective and nonlinear modeling capabilities derived from the differential equa
tion our method offers a concise solution to effectively bridge the optimization
 gap in a single step bypassing the lengthy fine-tuning process. The comprehensi
ve experiments on 24 supervised and self-supervised pre-trained models across 10
 downstream datasets demonstrate impressive performances and showcase its broad
adaptability even in low-data scenarios.
****
Video ReCap: Recursive Captioning of Hour-Long Videos
Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani,
Gedas Bertasius; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2024, pp. 18198-18208
Most video captioning models are designed to process short video clips of few se
conds and output text describing low-level visual concepts (e.g. objects scenes
atomic actions). However most real-world videos last for minutes or hours and ha
ve a complex hierarchical structure spanning different temporal granularities. W
e propose Video ReCap a recursive video captioning model that can process video
inputs of dramatically different lengths (from 1 second to 2 hours) and output v
ideo captions at multiple hierarchy levels. The recursive video-language archite
cture exploits the synergy between different video hierarchies and can process h
our-long videos efficiently. We utilize a curriculum learning training scheme to
 learn the hierarchical structure of videos starting from clip-level captions de
scribing atomic actions then focusing on segment-level descriptions and concludi
ng with generating summaries for hour-long videos. Furthermore we introduce Ego4
D-HCap dataset by augmenting Ego4D with 8267 manually collected long-range video
 summaries. Our recursive model can flexibly generate captions at different hier
archy levels while also being useful for other complex video understanding tasks
 such as VideoQA on EgoSchema. Data code and models are publicly available at ht
tps://sites.google.com/view/vidrecap.
****
Towards Realistic Scene Generation with LiDAR Diffusion Models
Haoxi Ran, Vitor Guizilini, Yue Wang; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14738-14748
Diffusion models (DMs) excel in photo-realistic image synthesis but their adapta
tion to LiDAR scene generation poses a substantial hurdle. This is primarily bec
ause DMs operating in the point space struggle to preserve the curve-like patter
ns and 3D geometry of LiDAR scenes which consumes much of their representation p
ower. In this paper we propose LiDAR Diffusion Models (LiDMs) to generate LiDAR-
realistic scenes from a latent space tailored to capture the realism of LiDAR sc
enes by incorporating geometric priors into the learning pipeline. Our method ta
rgets three major desiderata: pattern realism geometry realism and object realis
m. Specifically we introduce curve-wise compression to simulate real-world LiDAR
 patterns point-wise coordinate supervision to learn scene geometry and patch-wi
se encoding for a full 3D object context. With these three core designs our meth
od achieves competitive performance on unconditional LiDAR generation in 64-beam
 scenario and state of the art on conditional LiDAR generation while maintaining
 high efficiency compared to point-based DMs (up to 107xfaster). Furthermore by
compressing LiDAR scenes into a latent space we enable the controllability of DM
s with various conditions such as semantic maps camera views and text prompts. O
ur code and pretrained weights are available at https://github.com/hancyran/LiDA
R-Diffusion.
****
Diffusion Reflectance Map: Single-Image Stochastic Inverse Rendering of Illumina
tion and Reflectance
Yuto Enyo, Ko Nishino; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2024, pp. 11873-11883
Reflectance bounds the frequency spectrum of illumination in the object appearan
ce. In this paper we introduce the first stochastic inverse rendering method whi

ch recovers the attenuated frequency spectrum of an illumination jointly with the reflectance of an object of known geometry from a single image. Our key idea is to solve this blind inverse problem in the reflectance map an appearance representation invariant to the underlying geometry by learning to reverse the image formation with a novel diffusion model which we refer to as the Diffusion Reflectance Map Network (DRMNet). Given an observed reflectance map converted and completed from the single input image DRMNet generates a reflectance map corresponding to a perfect mirror sphere while jointly estimating the reflectance. The forward process can be understood as gradually filtering a natural illumination with lower and lower frequency reflectance and additive Gaussian noise. DRMNet learns to invert this process with two subnetworks IllNet and RefNet which work in concert towards this joint estimation. The network is trained on an extensive synthetic dataset and is demonstrated to generalize to real images showing state-of-the-art accuracy on established datasets.
****
Universal Segmentation at Arbitrary Granularity with Language Instruction
Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, Yansong Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3459-3469
This paper aims to achieve universal segmentation of arbitrary semantic level. Despite significant progress in recent years specialist segmentation approaches are limited to specific tasks and data distribution. Retraining a new model for adaptation to new scenarios or settings takes expensive computation and time cost which raises the demand for versatile and universal segmentation model that can cater to various granularity. Although some attempts have been made for unifying different segmentation tasks or generalization to various scenarios limitations in the definition of paradigms and input-output spaces make it difficult for them to achieve accurate understanding of content at arbitrary granularity. To this end we present UniLSeg a universal segmentation model that can perform segmentation at any semantic level with the guidance of language instructions. For training UniLSeg we reorganize a group of tasks from original diverse distributions into a unified data format where images with texts describing segmentation targets as input and corresponding masks are output. Combined with a automatic annotation engine for utilizing numerous unlabeled data UniLSeg achieves excellent performance on various tasks and settings surpassing both specialist and unified segmentation models.
****
GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians
Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20299-20309
We introduce GaussianAvatars a new method to create photorealistic head avatars that are fully controllable in terms of expression pose and viewpoint. The core idea is a dynamic 3D representation based on 3D Gaussian splats that are rigged to a parametric morphable face model. This combination facilitates photorealistic rendering while allowing for precise animation control via the underlying parametric model e.g. through expression transfer from a driving sequence or by manually changing the morphable model parameters. We parameterize each splat by a local coordinate frame of a triangle and optimize for explicit displacement offset to obtain a more accurate geometric representation. During avatar reconstruction we jointly optimize for the morphable model parameters and Gaussian splat parameters in an end-to-end fashion. We demonstrate the animation capabilities of our photorealistic avatar in several challenging scenarios. For instance we show reenactments from a driving video where our method outperforms existing works by a significant margin.
****
MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI
Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, R

enliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, Wenhu Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9556-9567
We introduce MMMU: a new benchmark designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. MMMU includes 11.5K meticulously collected multimodal questions from college exams quizzes and textbooks covering six core disciplines: Art & Design Business Science Health & Medicine Humanities & Social Science and Tech & Engineering. These questions span 30 subjects and 183 subfields comprising 30 highly heterogeneous image types such as charts diagrams maps tables music sheets and chemical structures. Unlike existing benchmarks MMMU focuses on advanced perception and reasoning with domain-specific knowledge challenging models to perform tasks akin to those faced by experts. The evaluation of 28 open-source LMMs as well as the proprietary GPT-4V(ision) and Gemini highlights the substantial challenges posed by MMMU. Even the advanced GPT-4V and Gemini Ultra only achieve accuracies of 56% and 59% respectively indicating significant room for improvement. We believe MMMU will stimulate the community to build next-generation multimodal foundation models towards expert artificial general intelligence.
****
Layout-Agnostic Scene Text Image Synthesis with Diffusion Models
Qilong Zhangli, Jindong Jiang, Di Liu, Licheng Yu, Xiaoliang Dai, Ankit Ramchandani, Guan Pang, Dimitris N. Metaxas, Praveen Krishnan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7496-7506
While diffusion models have significantly advanced the quality of image generation their capability to accurately and coherently render text within these images remains a substantial challenge. Conventional diffusion-based methods for scene text generation are typically limited by their reliance on an intermediate layout output. This dependency often results in a constrained diversity of text styles and fonts an inherent limitation stemming from the deterministic nature of the layout generation phase. To address these challenges this paper introduces SceneTextGen a novel diffusion-based model specifically designed to circumvent the need for a predefined layout stage. By doing so SceneTextGen facilitates a more natural and varied representation of text. The novelty of SceneTextGen lies in its integration of three key components: a character-level encoder for capturing detailed typographic properties coupled with a character-level instance segmentation model and a word-level spotting model to address the issues of unwanted text generation and minor character inaccuracies. We validate the performance of our method by demonstrating improved character recognition rates on generated images across different public visual text datasets in comparison to both standard diffusion based methods and text specific methods.
****
EarthLoc: Astronaut Photography Localization by Indexing Earth from Space
Gabriele Berton, Alex Stoken, Barbara Caputo, Carlo Masone; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12754-12764
Astronaut photography spanning six decades of human spaceflight presents a unique Earth observations dataset with immense value for both scientific research and disaster response. Despite its significance accurately localizing the geographical extent of these images crucial for effective utilization poses substantial challenges. Current manual localization efforts are time-consuming motivating the need for automated solutions. We propose a novel approach - leveraging image retrieval - to address this challenge efficiently. We introduce innovative training techniques including Year-Wise Data Augmentation and a Neutral-Aware Multi-Similarity Loss which contribute to the development of a high-performance model EarthLoc. We develop six evaluation datasets and perform a comprehensive benchmark comparing EarthLoc to existing methods showcasing its superior efficiency and accuracy. Our approach marks a significant advancement in automating the localization of astronaut photography which will help bridge a critical gap in Earth observations data. Code and datasets are available at this https://github.com/gmbert

on/EarthLoc
****
SmartMask: Context Aware High-Fidelity Mask Generation for Fine-grained Object I
nsertion and Layout Control
Jaskirat Singh, Jianming Zhang, Qing Liu, Cameron Smith, Zhe Lin, Liang Zheng; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 6497-6506
The field of generative image inpainting and object insertion has made significa
nt progress with the recent advent of latent diffusion models. Utilizing a preci
se object mask can greatly enhance these applications. However due to the challe
nges users encounter in creating high-fidelity masks there is a tendency for the
se methods to rely on more coarse masks (e.g. bounding box) for these applicatio
ns. This results in limited control and compromised background content preservat
ion. To overcome these limitations we introduce SmartMask which allows any novic
e user to create detailed masks for precise object insertion. Combined with a Co
ntrolNet-Inpaint model our experiments demonstrate that SmartMask achieves super
ior object insertion quality preserving the background content more effectively
than previous methods. Notably unlike prior works the proposed approach can also
 be used even without user-mask guidance which allows it to perform mask-free ob
ject insertion at diverse positions and scales. Furthermore we find that when us
ed iteratively with a novel instruction-tuning based planning model SmartMask ca
n be used to design detailed layouts from scratch. As compared with user-scribbl
e based layout design we observe that SmartMask allows for better quality output
s with layout-to-image generation methods.
****
Text-Image Alignment for Diffusion-Based Perception
Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogerio Guimaraes, Pietro Perona
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2024, pp. 13883-13893
Diffusion models are generative models with impressive text-to-image synthesis c
apabilities and have spurred a new wave of creative methods for classical machin
e learning tasks. However the best way to harness the perceptual knowledge of th
ese generative models for visual tasks is still an open question. Specifically i
t is unclear how to use the prompting interface when applying diffusion backbone
s to vision tasks. We find that automatically generated captions can improve tex
t-image alignment and significantly enhance a model's cross-attention maps leadi
ng to better perceptual performance. Our approach improves upon the current stat
e-of-the-art in diffusion-based semantic segmentation on ADE20K and the current
overall SOTA for depth estimation on NYUv2. Furthermore our method generalizes t
o the cross-domain setting. We use model personalization and caption modificatio
ns to align our model to the target domain and find improvements over unaligned
baselines. Our cross-domain object detection model trained on Pascal VOC achieve
s SOTA results on Watercolor2K. Our cross-domain segmentation method trained on
Cityscapes achieves SOTA results on Dark Zurich-val and Nighttime Driving. Proje
ct page: vision.caltech.edu/TADP/. Code: github.com/damaggu/TADP
****
Customization Assistant for Text-to-Image Generation
Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, Tong Sun; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9182-9191
Customizing pre-trained text-to-image generation model has attracted massive res
earch interest recently due to its huge potential in real-world applications. Al
though existing methods are able to generate creative content for a novel concep
t contained in single user-input image their capability are still far from perfe
ction. Specifically most existing methods require fine-tuning the generative mod
el on testing images. Some existing methods do not require fine-tuning while the
ir performance are unsatisfactory. Furthermore the interaction between users and
 models are still limited to directive and descriptive prompts such as instructi
ons and captions. In this work we build a customization assistant based on pre-t
rained large language model and diffusion model which can not only perform custo
mized generation in a tuning-free manner but also enable more user-friendly inte

ractions: users can chat with the assistant and input either ambiguous text or clear instruction. Specifically we propose a new framework consists of a new model design and a novel training strategy. The resulting assistant can perform customized generation in 2-5 seconds without any test time fine-tuning. Extensive experiments are conducted competitive results have been obtained across different domains illustrating the effectiveness of the proposed method.
****
GaussianEditor: Editing 3D Gaussians Delicately with Text Instructions
Junjie Wang, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20902-20911
Recently impressive results have been achieved in 3D scene editing with text instructions based on a 2D diffusion model. However current diffusion models primarily generate images by predicting noise in the latent space and the editing is usually applied to the whole image which makes it challenging to perform delicate especially localized editing for 3D scenes. Inspired by recent 3D Gaussian splatting we propose a systematic framework named GaussianEditor to edit 3D scenes delicately via 3D Gaussians with text instructions. Benefiting from the explicit property of 3D Gaussians we design a series of techniques to achieve delicate editing. Specifically we first extract the region of interest (RoI) corresponding to the text instruction aligning it to 3D Gaussians. The Gaussian RoI is further used to control the editing process. Our framework can achieve more delicate and precise editing of 3D scenes than previous methods while enjoying much faster training speed i.e. within 20 minutes on a single V100 GPU more than twice as fast as Instruct-NeRF2NeRF (45 minutes -- 2 hours). The project page is at GaussianEditor.github.io.
****
MemFlow: Optical Flow Estimation and Prediction with Memory
Qiaole Dong, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19068-19078
Optical flow is a classical task that is important to the vision community. Classical optical flow estimation uses two frames as input whilst some recent methods consider multiple frames to explicitly model long-range information. The former ones limit their ability to fully leverage temporal coherence along the video sequence; and the latter ones incur heavy computational overhead typically not possible for real-time flow estimation. Some multi-frame-based approaches even necessitate unseen future frames for current estimation compromising real-time applicability in safety-critical scenarios. To this end we present MemFlow a real-time method for optical flow estimation and prediction with memory. Our method enables memory read-out and update modules for aggregating historical motion information in real-time. Furthermore we integrate resolution-adaptive re-scaling to accommodate diverse video resolutions. Besides our approach seamlessly extends to the future prediction of optical flow based on past observations. Leveraging effective historical motion aggregation our method outperforms VideoFlow with fewer parameters and faster inference speed on Sintel and KITTI-15 datasets in terms of generalization performance. At the time of submission MemFlow also leads in performance on the 1080p Spring dataset. Codes and models will be available at: https://dqiaole.github.io/MemFlow/.
****
Novel Class Discovery for Ultra-Fine-Grained Visual Categorization
Yu Liu, Yaqi Cai, Qi Jia, Binglin Qiu, Weimin Wang, Nan Pu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17679-17688
Ultra-fine-grained visual categorization (Ultra-FGVC) aims at distinguishing highly similar sub-categories within fine-grained objects such as different soybean cultivars. Compared to traditional fine-grained visual categorization Ultra-FGVC encounters more hurdles due to the small inter-class and large intra-class variation. Given these challenges relying on human annotation for Ultra-FGVC is impractical. To this end our work introduces a novel task termed Ultra-Fine-Grained Novel Class Discovery (UFG-NCD) which leverages partially annotated data to ide

ntify new categories of unlabeled images for Ultra-FGVC. To tackle this problem we devise a Region-Aligned Proxy Learning (RAPL) framework which comprises a Channel-wise Region Alignment (CRA) module and a Semi-Supervised Proxy Learning (SemiPL) strategy. The CRA module is designed to extract and utilize discriminative features from local regions facilitating knowledge transfer from labeled to unlabeled classes. Furthermore SemiPL strengthens representation learning and knowledge transfer with proxy-guided supervised learning and proxy-guided contrastive learning. Such techniques leverage class distribution information in the embedding space improving the mining of subtle differences between labeled and unlabeled ultra-fine-grained classes. Extensive experiments demonstrate that RAPL significantly outperforms baselines across various datasets indicating its effectiveness in handling the challenges of UFG-NCD. Code is available at https://github.com/SSDUT-Caiyq/UFG-NCD.
****
GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos
Tomáš Sou?ek, Dima Damen, Michael Wray, Ivan Laptev, Josef Sivic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6561-6571
We address the task of generating temporally consistent and physically plausible images of actions and object state transformations. Given an input image and a text prompt describing the targeted transformation our generated images preserve the environment and transform objects in the initial image. Our contributions are threefold. First we leverage a large body of instructional videos and automatically mine a dataset of triplets of consecutive frames corresponding to initial object states actions and resulting object transformations. Second equipped with this data we develop and train a conditioned diffusion model dubbed GenHowTo. Third we evaluate GenHowTo on a variety of objects and actions and show superior performance compared to existing methods. In particular we introduce a quantitative evaluation where GenHowTo achieves 88% and 74% on seen and unseen interaction categories respectively outperforming prior work by a large margin.
****
Paint-it: Text-to-Texture Synthesis via Deep Convolutional Texture Map Optimization and Physically-Based Rendering
Kim Youwang, Tae-Hyun Oh, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4347-4356
We present Paint-it a text-driven high-fidelity texture map synthesis method for 3D meshes via neural re-parameterized texture optimization. Paint-it synthesizes texture maps from a text description by synthesis-through-optimization exploiting the Score-Distillation Sampling (SDS). We observe that directly applying SDS yields undesirable texture quality due to its noisy gradients. We reveal the importance of texture parameterization when using SDS. Specifically we propose Deep Convolutional Physically-Based Rendering (DC-PBR) parameterization which re-parameterizes the physically-based rendering (PBR) texture maps with randomly initialized convolution-based neural kernels instead of a standard pixel-based parameterization. We show that DC-PBR inherently schedules the optimization curriculum according to texture frequency and naturally filters out the noisy signals from SDS. In experiments Paint-it obtains remarkable quality PBR texture maps within 15 min. given only a text description. We demonstrate the generalizability and practicality of Paint-it by synthesizing high-quality texture maps for large-scale mesh datasets and showing test-time applications such as relighting and material control using a popular graphics engine.
****
HiKER-SGG: Hierarchical Knowledge Enhanced Robust Scene Graph Generation
Ce Zhang, Simon Stepputtis, Joseph Campbell, Katia Sycara, Yaqi Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28233-28243
Being able to understand visual scenes is a precursor for many downstream tasks including autonomous driving robotics and other vision-based approaches. A common approach enabling the ability to reason over visual data is Scene Graph Genera

tion (SGG); however many existing approaches assume undisturbed vision i.e. the absence of real-world corruptions such as fog snow smoke as well as non-uniform perturbations like sun glare or water drops. In this work we propose a novel SGG benchmark containing procedurally generated weather corruptions and other transformations over the Visual Genome dataset. Further we introduce a corresponding approach Hierarchical Knowledge Enhanced Robust Scene Graph Generation (HiKER-SGG) providing a strong baseline for scene graph generation under such challenging setting. At its core HiKER-SGG utilizes a hierarchical knowledge graph in order to refine its predictions from coarse initial estimates to detailed predictions. In our extensive experiments we show that HiKER-SGG does not only demonstrate superior performance on corrupted images in a zero-shot manner but also outperforms current state-of-the-art methods on uncorrupted SGG tasks. Code is available at https://github.com/zhangce01/HiKER-SGG.
****
DiffusionGAN3D: Boosting Text-guided 3D Generation and Domain Adaptation by Combining 3D GANs and Diffusion Priors
Biwen Lei, Kai Yu, Mengyang Feng, Miaomiao Cui, Xuansong Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10487-10497
Text-guided domain adaptation and generation of 3D-aware portraits find many applications in various fields. However due to the lack of training data and the challenges in handling the high variety of geometry and appearance the existing methods for these tasks suffer from issues like inflexibility instability and low fidelity. In this paper we propose a novel framework DiffusionGAN3D which boosts text-guided 3D domain adaptation and generation by combining 3D GANs and diffusion priors. Specifically we integrate the pre-trained 3D generative models (e.g. EG3D) and text-to-image diffusion models. The former provides a strong foundation for stable and high-quality avatar generation from text. And the diffusion models in turn offer powerful priors and guide the 3D generator finetuning with informative direction to achieve flexible and efficient text-guided domain adaptation. To enhance the diversity in domain adaptation and the generation capability in text-to-avatar we introduce the relative distance loss and case-specific learnable triplane respectively. Besides we design a progressive texture refinement module to improve the texture quality for both tasks above. Extensive experiments demonstrate that the proposed framework achieves excellent results in both domain adaptation and text-to-avatar tasks outperforming existing methods in terms of generation quality and efficiency. The project homepage is at https://younglbw.github.io/DiffusionGAN3D-homepage/.
****
Physics-Aware Hand-Object Interaction Denoising
Haowen Luo, Yunze Liu, Li Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2341-2350
The credibility and practicality of a reconstructed hand-object interaction sequence depend largely on its physical plausibility. However due to high occlusions during hand-object interaction physical plausibility remains a challenging criterion for purely vision-based tracking methods. To address this issue and enhance the results of existing hand trackers this paper proposes a novel physically-aware hand motion de-noising method. Specifically we introduce two learned loss terms that explicitly capture two crucial aspects of physical plausibility: grasp credibility and manipulation feasibility. These terms are used to train a physically-aware de-noising network. Qualitative and quantitative experiments demonstrate that our approach significantly improves both fine-grained physical plausibility and overall pose accuracy surpassing current state-of-the-art de-noising methods.
****
VastGaussian: Vast 3D Gaussians for Large Scene Reconstruction
Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, Wenming Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5166-5175

Existing NeRF-based methods for large scene reconstruction often have limitation
s in visual quality and rendering speed. While the recent 3D Gaussian Splatting
works well on small-scale and object-centric scenes scaling it up to large scene
s poses challenges due to limited video memory long optimization time and notice
able appearance variations. To address these challenges we present VastGaussian
the first method for high-quality reconstruction and real-time rendering on larg
e scenes based on 3D Gaussian Splatting. We propose a progressive partitioning s
trategy to divide a large scene into multiple cells where the training cameras a
nd point cloud are properly distributed with an airspace-aware visibility criter
ion. These cells are merged into a complete scene after parallel optimization. W
e also introduce decoupled appearance modeling into the optimization process to
reduce appearance variations in the rendered images. Our approach outperforms ex
isting NeRF-based methods and achieves state-of-the-art results on multiple larg
e scene datasets enabling fast optimization and high-fidelity real-time renderin
g.
****
Edit One for All: Interactive Batch Image Editing
Thao Nguyen, Utkarsh Ojha, Yuheng Li, Haotian Liu, Yong Jae Lee; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 8271-8280
In recent years image editing has advanced remarkably. With increased human cont
rol it is now possible to edit an image in a plethora of ways; from specifying i
n text what we want to change to straight up dragging the contents of the image
in an interactive point-based manner. However most of the focus has remained on
editing single images at a time. Whether and how we can simultaneously edit larg
e batches of images has remained understudied. With the goal of minimizing human
supervision in the editing process this paper presents a novel method for inter
active batch image editing using StyleGAN as the medium. Given an edit specified
by users in an example image (e.g. make the face frontal) our method can automa
tically transfer that edit to other test images so that regardless of their init
ial state (pose) they all arrive at the same final state (e.g. all facing front)
. Extensive experiments demonstrate that edits performed using our method have s
imilar visual quality to existing single-image-editing methods while having more
visual consistency and saving significant time and human effort.
****
Rethinking Boundary Discontinuity Problem for Oriented Object Detection
Hang Xu, Xinyuan Liu, Haonan Xu, Yike Ma, Zunjie Zhu, Chenggang Yan, Feng Dai; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 17406-17415
Oriented object detection has been developed rapidly in the past few years where
rotation equivariance is crucial for detectors to predict rotated boxes. It is
expected that the prediction can maintain the corresponding rotation when object
s rotate but severe mutation in angular prediction is sometimes observed when ob
jects rotate near the boundary angle which is well-known boundary discontinuity
problem. The problem has been long believed to be caused by the sharp loss incre
ase at the angular boundary and widely used joint-optim IoU-like methods deal wi
th this problem by loss-smoothing. However we experimentally find that even stat
e-of-the-art IoU-like methods actually fail to solve the problem. On further ana
lysis we find that the key to solution lies in encoding mode of the smoothing fu
nction rather than in joint or independent optimization. In existing IoU-like me
thods the model essentially attempts to fit the angular relationship between box
and object where the break point at angular boundary makes the predictions high
ly unstable. To deal with this issue we propose a dual-optimization paradigm for
angles. We decouple reversibility and joint-optim from single smoothing functio
n into two distinct entities which for the first time achieves the objectives of
both correcting angular boundary and blending angle with other parameters. Exte
nsive experiments on multiple datasets show that boundary discontinuity problem
is well-addressed. Moreover typical IoU-like methods are improved to the same le
vel without obvious performance gap. The code is available at https://github.com
/hangxu-cv/cvpr24acm.

****
Deformable One-shot Face Stylization via DINO Semantic Guidance
Yang Zhou, Zichong Chen, Hui Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7787-7796
This paper addresses the complex issue of one-shot face stylization focusing on the simultaneous consideration of appearance and structure where previous methods have fallen short. We explore deformation-aware face stylization that diverges from traditional single-image style reference opting for a real-style image pair instead. The cornerstone of our method is the utilization of a self-supervised vision transformer specifically DINO-ViT to establish a robust and consistent facial structure representation across both real and style domains. Our stylization process begins by adapting the StyleGAN generator to be deformation-aware through the integration of spatial transformers (STN). We then introduce two innovative constraints for generator fine-tuning under the guidance of DINO semantics: i) a directional deformation loss that regulates directional vectors in DINO space and ii) a relative structural consistency constraint based on DINO token self-similarities ensuring diverse generation. Additionally style-mixing is employed to align the color generation with the reference minimizing inconsistent correspondences. This framework delivers enhanced deformability for general one-shot face stylization achieving notable efficiency with a fine-tuning duration of approximately 10 minutes. Extensive qualitative and quantitative comparisons demonstrate our superiority over state-of-the-art one-shot face stylization methods. Code is available at https://github.com/zichongc/DoesFS
****
SleepVST: Sleep Staging from Near-Infrared Video Signals using Pre-Trained Transformers
Jonathan F. Carter, João Jorge, Oliver Gibson, Lionel Tarassenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12479-12489
Advances in camera-based physiological monitoring have enabled the robust non-contact measurement of respiration and the cardiac pulse which are known to be indicative of the sleep stage. This has led to research into camera-based sleep monitoring as a promising alternative to "gold-standard" polysomnography which is cumbersome expensive to administer and hence unsuitable for longer-term clinical studies. In this paper we introduce SleepVST a transformer model which enables state-of-the-art performance in camera-based sleep stage classification (sleep staging). After pre-training on contact sensor data SleepVST outperforms existing methods for cardio-respiratory sleep staging on the SHHS and MESA datasets achieving total Cohen's kappa scores of 0.75 and 0.77 respectively. We then show that SleepVST can be successfully transferred to cardio-respiratory waveforms extracted from video enabling fully contact-free sleep staging. Using a video dataset of 50 nights we achieve a total accuracy of 78.8% and a Cohen's \kappa of 0.71 in four-class video-based sleep staging setting a new state-of-the-art in the domain.
****
Coarse-to-Fine Latent Diffusion for Pose-Guided Person Image Synthesis
Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, Jianhuang Lai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6420-6429
Diffusion model is a promising approach to image generation and has been employed for Pose-Guided Person Image Synthesis (PGPIS) with competitive performance. While existing methods simply align the person appearance to the target pose they are prone to overfitting due to the lack of a high-level semantic understanding on the source person image. In this paper we propose a novel Coarse-to-Fine Latent Diffusion (CFLD) method for PGPIS. In the absence of image-caption pairs and textual prompts we develop a novel training paradigm purely based on images to control the generation process of a pre-trained text-to-image diffusion model. A perception-refined decoder is designed to progressively refine a set of learnable queries and extract semantic understanding of person images as a coarse-grained prompt. This allows for the decoupling of fine-grained appearance and pose in

formation controls at different stages and thus circumventing the potential over fitting problem. To generate more realistic texture details a hybrid-granularity attention module is proposed to encode multi-scale fine-grained appearance features as bias terms to augment the coarse-grained prompt. Both quantitative and qualitative experimental results on the DeepFashion benchmark demonstrate the superiority of our method over the state of the arts for PGPIS. Code is available at https://github.com/YanzuoLu/CFLD.
****
Watermark-embedded Adversarial Examples for Copyright Protection against Diffusion Models
Peifei Zhu, Tsubasa Takahashi, Hirokatsu Kataoka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24420-24430
Diffusion Models (DMs) have shown remarkable capabilities in various image-generation tasks. However there are growing concerns that DMs could be used to imitate unauthorized creations and thus raise copyright issues. To address this issue we propose a novel framework that embeds personal watermarks in the generation of adversarial examples. Such examples can force DMs to generate images with visible watermarks and prevent DMs from imitating unauthorized images. We construct a generator based on conditional adversarial networks and design three losses (adversarial loss GAN loss and perturbation loss) to generate adversarial examples that have subtle perturbation but can effectively attack DMs to prevent copyright violations. Training a generator for a personal watermark by our method only requires 5-10 samples within 2-3 minutes and once the generator is trained it can generate adversarial examples with that watermark significantly fast (0.2s per image). We conduct extensive experiments in various conditional image-generation scenarios. Compared to existing methods that generate images with chaotic textures our method adds visible watermarks on the generated images which is a more straightforward way to indicate copyright violations. We also observe that our adversarial examples exhibit good transferability across unknown generative models. Therefore this work provides a simple yet powerful way to protect copyright from DM-based imitation.
****
TCP:Textual-based Class-aware Prompt tuning for Visual-Language Model
Hantao Yao, Rui Zhang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23438-23448
Prompt tuning represents a valuable technique for adapting pre-trained visual-language models (VLM) to various downstream tasks. Recent advancements in CoOp-based methods propose a set of learnable domain-shared or image-conditional textual tokens to facilitate the generation of task-specific textual classifiers. However those textual tokens have a limited generalization ability regarding unseen domains as they cannot dynamically adjust to the distribution of testing classes. To tackle this issue we present a novel Textual-based Class-aware Prompt tuning (TCP) that explicitly incorporates prior knowledge about classes to enhance their discriminability. The critical concept of TCP involves leveraging Textual Knowledge Embedding (TKE) to map the high generalizability of class-level textual knowledge into class aware textual tokens. By seamlessly integrating these class-aware prompts into the Text Encoder a dynamic class-aware classifier is generated to enhance discriminability for unseen domains. During inference TKE dynamically generates class-aware prompts related to the unseen classes. Comprehensive evaluations demonstrate that TKE serves as a plug-and-play module effortlessly combinable with existing methods. Furthermore TCP consistently achieves superior performance while demanding less training time.
****
OMG: Towards Open-vocabulary Motion Generation via Mixture of Controllers
Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibei Yang, Xin Chen, Jingyi Yu, Lan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 482-493
We have recently seen tremendous progress in realistic text-to-motion generation. Yet the existing methods often fail or produce implausible motions with unseen

text inputs which limits the applications. In this paper we present OMG a novel
 framework which enables compelling motion generation from zero-shot open-vocabu
lary text prompts. Our key idea is to carefully tailor the pretrain-then-finetun
e paradigm into the text-to-motion generation. At the pre-training stage our mod
el improves the generation ability by learning the rich out-of-domain inherent m
otion traits. To this end we scale up a large unconditional diffusion model up t
o 1B parameters so as to utilize the massive unlabeled motion data up to over 20
M motion instances. At the subsequent fine-tuning stage we introduce motion Cont
rolNet which incorporates text prompts as conditioning information through a tra
inable copy of the pre-trained model and the proposed novel Mixture-of-Controlle
rs (MoC) block. MoC block adaptively recognizes various ranges of the sub-motion
s with a cross-attention mechanism and processes them separately with the text-t
oken-specific experts. Such a design effectively aligns the CLIP token embedding
s of text prompts to various ranges of compact and expressive motion features. E
xtensive experiments demonstrate that our OMG achieves significant improvements
over the state-of-the-art methods on zero-shot text-to-motion generation. Projec
t page: https://tr3e.github.io/omg-page.
****
TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Unders
tanding
Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, Lu Hou; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14313-1
4323
This work proposes TimeChat a time-sensitive multimodal large language model spe
cifically designed for long video understanding. Our model incorporates two key
architectural contributions: (1) a timestamp-aware frame encoder that binds visu
al content with the timestamp of each frame and (2) a sliding video Q-Former tha
t produces a video token sequence of varying lengths to accommodate videos of va
rious durations. Additionally we construct an instruction-tuning dataset encompa
ssing 6 tasks and a total of 125K instances to further enhance TimeChat's instru
ction-following performance. Experiment results across various video understandi
ng tasks such as dense captioning temporal grounding and highlight detection dem
onstrate TimeChat's strong zero-shot temporal localization and reasoning capabil
ities. For example it achieves +9.2 F1 score and +2.8 CIDEr on YouCook2 +5.8 HIT
@1 on QVHighlights and +27.5 R@1 (IoU=0.5) on Charades-STA compared to state-of-
the-art video large language models holding the potential to serve as a versatil
e video assistant for long-form video comprehension tasks and satisfy realistic
user requirements.
****
Align Your Gaussians: Text-to-4D with Dynamic 3D Gaussians and Composed Diffusio
n Models
Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, Karsten Kreis; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2024, pp. 8576-8588
Text-guided diffusion models have revolutionized image and video generation and
have also been successfully used for optimization-based 3D object synthesis. Her
e we instead focus on the underexplored text-to-4D setting and synthesize dynami
c animated 3D objects using score distillation methods with an additional tempor
al dimension. Compared to previous work we pursue a novel compositional generati
on-based approach and combine text-to-image text-to-video and 3D-aware multiview
 diffusion models to provide feedback during 4D object optimization thereby simu
ltaneously enforcing temporal consistency high-quality visual appearance and rea
listic geometry. Our method called Align Your Gaussians (AYG) leverages dynamic
3D Gaussian Splatting with deformation fields as 4D representation. Crucial to A
YG is a novel method to regularize the distribution of the moving 3D Gaussians a
nd thereby stabilize the optimization and induce motion. We also propose a motio
n amplification mechanism as well as a new autoregressive synthesis scheme to ge
nerate and combine multiple 4D sequences for longer generation. These techniques
 allow us to synthesize vivid dynamic scenes outperform previous work qualitativ
ely and quantitatively and achieve state-of-the-art text-to-4D performance. Due

to the Gaussian 4D representation different 4D animations can be seamlessly comb ined as we demonstrate. AYG opens up promising avenues for animation simulation and digital content creation as well as synthetic data generation.
****
PDF: A Probability-Driven Framework for Open World 3D Point Cloud Semantic Segme ntation
Jinfeng Xu, Siyuan Yang, Xianzhi Li, Yuan Tang, Yixue Hao, Long Hu, Min Chen; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5977-5986
Existing point cloud semantic segmentation networks cannot identify unknown clas ses and update their knowledge due to a closed-set and static perspective of the real world which would induce the intelligent agent to make bad decisions. To a ddress this problem we propose a Probability-Driven Framework (PDF) for open wor ld semantic segmentation that includes (i) a lightweight U-decoder branch to ide ntify unknown classes by estimating the uncertainties (ii) a flexible pseudo-lab eling scheme to supply geometry features along with probability distribution fea tures of unknown classes by generating pseudo labels and (iii) an incremental kn owledge distillation strategy to incorporate novel classes into the existing kno wledge base gradually. Our framework enables the model to behave like human bein gs which could recognize unknown objects and incrementally learn them with the c orresponding knowledge. Experimental results on the S3DIS and ScanNetv2 datasets demonstrate that the proposed PDF outperforms other methods by a large margin i n both important tasks of open world semantic segmentation.
****
Test-Time Domain Generalization for Face Anti-Spoofing
Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, Lizhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2024, pp. 175-187
Face Anti-Spoofing (FAS) is pivotal in safeguarding facial recognition systems a gainst presentation attacks. While domain generalization (DG) methods have been developed to enhance FAS performance they predominantly focus on learning domain -invariant features during training which may not guarantee generalizability to unseen data that differs largely from the source distributions. Our insight is t hat testing data can serve as a valuable resource to enhance the generalizabilit y beyond mere evaluation for DG FAS. In this paper we introduce a novel Test-Tim e Domain Generalization (TTDG) framework for FAS which leverages the testing dat a to boost the model's generalizability. Our method consisting of Test-Time Styl e Projection (TTSP) and Diverse Style Shifts Simulation (DSSS) effectively proje cts the unseen data to the seen domain space. In particular we first introduce t he innovative TTSP to project the styles of the arbitrarily unseen samples of th e testing distribution to the known source space of the training distributions. We then design the efficient DSSS to synthesize diverse style shifts via learnab le style bases with two specifically designed losses in a hyperspherical feature space. Our method eliminates the need for model updates at the test time and ca n be seamlessly integrated into not only the CNN but also ViT backbones. Compreh ensive experiments on widely used cross-domain FAS benchmarks demonstrate our me thod's state-of-the-art performance and effectiveness.
****
DiffusionMTL: Learning Multi-Task Denoising Diffusion Model from Partially Annot ated Data
Hanrong Ye, Dan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2024, pp. 27960-27969
Recently there has been an increased interest in the practical problem of learni ng multiple dense scene understanding tasks from partially annotated data where each training sample is only labeled for a subset of the tasks. The missing of t ask labels in training leads to low-quality and noisy predictions as can be obse rved from state-of-the-art methods. To tackle this issue we reformulate the part ially-labeled multi-task dense prediction as a pixel-level denoising problem and propose a novel multi-task denoising diffusion framework coined as DiffusionMTL . It designs a joint diffusion and denoising paradigm to model a potential noisy

distribution in the task prediction or feature maps and generate rectified outputs for different tasks. To exploit multi-task consistency in denoising we further introduce a Multi-Task Conditioning strategy which can implicitly utilize the complementary nature of the tasks to help learn the unlabeled tasks leading to an improvement in the denoising performance of the different tasks. Extensive quantitative and qualitative experiments demonstrate that the proposed multi-task denoising diffusion model can significantly improve multi-task prediction maps and outperform the state-of-the-art methods on three challenging multi-task benchmarks under two different partial-labeling evaluation settings. The code is available at https://prismformore.github.io/diffusionmtl/.
****
Spike-guided Motion Deblurring with Unknown Modal Spatiotemporal Alignment
Jiyuan Zhang, Shiyan Chen, Yajing Zheng, Zhaofei Yu, Tiejun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25047-25057
The traditional frame-based cameras that rely on exposure windows for imaging experience motion blur in high-speed scenarios. Frame-based deblurring methods lack reliable motion cues to restore sharp images under extreme blur conditions. The spike camera is a novel neuromorphic visual sensor that outputs spike streams with ultra-high temporal resolution. It can supplement the temporal information lost in traditional cameras and guide motion deblurring. However in real-world scenarios aligning discrete RGB images and continuous spike streams along both temporal and spatial axes is challenging due to the complexity of calibrating their coordinates device displacements in vibrations and time deviations. Misalignment of pixels leads to severe degradation of deblurring. We introduce the first framework for spike-guided motion deblurring without knowing the spatiotemporal alignment between spikes and images. To address the problem we first propose a novel three-stage network containing a basic deblurring net a carefully designed bi-directional deformable aligning module and a flow-based multi-scale fusion net. Experimental results demonstrate that our approach can effectively guide the image deblurring with unknown alignment surpassing the performance of other methods. Public project page: https://github.com/Leozhangjiyuan/UaSDN.
****
VRP-SAM: SAM with Visual Reference Prompt
Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, Zechao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23565-23574
In this paper we propose a novel Visual Reference Prompt (VRP) encoder that empowers the Segment Anything Model (SAM) to utilize annotated reference images as prompts for segmentation creating the VRP-SAM model. In essence VRP-SAM can utilize annotated reference images to comprehend specific objects and perform segmentation of specific objects in target image. It is note that the VRP encoder can support a variety of annotation formats for reference images including point box scribble and mask. VRP-SAM achieves a breakthrough within the SAM framework by extending its versatility and applicability while preserving SAM's inherent strengths thus enhancing user-friendliness. To enhance the generalization ability of VRP-SAM the VRP encoder adopts a meta-learning strategy. To validate the effectiveness of VRP-SAM we conducted extensive empirical studies on the Pascal and COCO datasets. Remarkably VRP-SAM achieved state-of-the-art performance in visual reference segmentation with minimal learnable parameters. Furthermore VRP-SAM demonstrates strong generalization capabilities allowing it to perform segmentation of unseen objects and enabling cross-domain segmentation. The source code and models will be available at https://github.com/syp2ysy/VRP-SAM
****
Discriminability-Driven Channel Selection for Out-of-Distribution Detection
Yue Yuan, Rundong He, Yicong Dong, Zhongyi Han, Yilong Yin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26171-26180
Out-of-distribution (OOD) detection is essential for deploying machine learning models in open-world environments. Activation-based methods are a key approach i

n OOD detection working to mitigate overconfident predictions of OOD data. These techniques rectifying anomalous activations enhancing the distinguishability between in-distribution (ID) data and OOD data. However they assume by default that every channel is necessary for OOD detection and rectify anomalous activations in each channel. Empirical evidence has shown that there is a significant difference among various channels in OOD detection and discarding some channels can greatly enhance the performance of OOD detection. Based on this insight we propose \underline D iscriminability-\underline D riven \underline C hannel \underline S election (DDCS) which leverages an adaptive channel selection by estimating the discriminative score of each channel to boost OOD detection. The discriminative score takes inter-class similarity and inter-class variance of training data into account. However the estimation of discriminative score itself is susceptible to anomalous activations. To better estimate score we pre-rectify anomalous activations for each channel mildly. The experimental results show that DDCS achieves state-of-the-art performance on CIFAR and ImageNet-1K benchmarks. Moreover DDCS can generalize to different backbones and OOD scores.
****
ManiFPT: Defining and Analyzing Fingerprints of Generative Models
Hae Jin Song, Mahyar Khayatkhoei, Wael AbdAlmageed; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10791-10801
Recent works have shown that generative models leave traces of their underlying generative process on the generated samples broadly referred to as fingerprints of a generative model and have studied their utility in detecting synthetic images from real ones. However the extend to which these fingerprints can distinguish between various types of synthetic image and help identify the underlying generative process remain under-explored. In particular the very definition of a fingerprint remains unclear to our knowledge. To that end in this work we formalize the definition of artifact and fingerprint in generative models propose an algorithm for computing them in practice and finally study its effectiveness in distinguishing a large array of different generative models. We find that using our proposed definition can significantly improve the performance on the task of identifying the underlying generative process from samples (model attribution) compared to existing methods. Additionally we study the structure of the fingerprints and observe that it is very predictive of the effect of different design choices on the generative process.
****
Real-time 3D-aware Portrait Video Relighting
Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, Lin Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6221-6231
Synthesizing realistic videos of talking faces under custom lighting conditions and viewing angles benefits various downstream applications like video conferencing. However most existing relighting methods are either time-consuming or unable to adjust the viewpoints. In this paper we present the first real-time 3D-aware method for relighting in-the-wild videos of talking faces based on Neural Radiance Fields (NeRF). Given an input portrait video our method can synthesize talking faces under both novel views and novel lighting conditions with a photo-realistic and disentangled 3D representation. Specifically we infer an albedo tri-plane as well as a shading tri-plane based on a desired lighting condition for each video frame with fast dual-encoders. We also leverage a temporal consistency network to ensure smooth transitions and reduce flickering artifacts. Our method runs at 32.98 fps on consumer-level hardware and achieves state-of-the-art results in terms of reconstruction quality lighting error lighting instability temporal consistency and inference speed. We demonstrate the effectiveness and interactivity of our method on various portrait videos with diverse lighting and viewing conditions.
****
3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting
Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, Siyu Tang; Proceedi

ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5020-5030
We introduce an approach that creates animatable human avatars from monocular videos using 3D Gaussian Splatting (3DGS). Existing methods based on neural radiance fields (NeRFs) achieve high-quality novel-view/novel-pose image synthesis but often require days of training and are extremely slow at inference time. Recently the community has explored fast grid structures for efficient training of clothed avatars. Albeit being extremely fast at training these methods can barely achieve an interactive rendering frame rate with around 15 FPS. In this paper we use 3D Gaussian Splatting and learn a non-rigid deformation network to reconstruct animatable clothed human avatars that can be trained within 30 minutes and rendered at real-time frame rates (50+ FPS). Given the explicit nature of our representation we further introduce as-isometric-as-possible regularizations on both the Gaussian mean vectors and the covariance matrices enhancing the generalization of our model on highly articulated unseen poses. Experimental results show that our method achieves comparable and even better performance compared to state-of-the-art approaches on animatable avatar creation from a monocular input while being 400x and 250x faster in training and inference respectively.
****
Quilt-LLaVA: Visual Instruction Tuning by Extracting Localized Narratives from Open-Source Histopathology Videos
Mehmet Saygin Seyfioglu, Wisdom O. Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, Linda Shapiro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13183-13192
Diagnosis in histopathology requires a global whole slide images (WSIs) analysis requiring pathologists to compound evidence from different WSI patches. The gigapixel scale of WSIs poses a challenge for histopathology multi-modal models. Training multi-model models for histopathology requires instruction tuning datasets which currently contain information for individual image patches without a spatial grounding of the concepts within each patch and without a wider view of the WSI. To bridge this gap we introduce QUILT-INSTRUCT a large-scale dataset of 107131 histopathology-specific instruction question/answer pairs grounded within diagnostically relevant image patches that make up the WSI. Our dataset is collected by leveraging educational histopathology videos from YouTube which provides spatial localization of narrations by automatically extracting the narrators' cursor positions. QUILT-INSTRUCT supports contextual reasoning by extracting diagnosis and supporting facts from the entire WSI. Using QUILT-INSTRUCT we train QUILT-LLAVA which can reason beyond the given single image patch enabling diagnostic reasoning across patches. To evaluate QUILT-LLAVA we propose a comprehensive evaluation dataset created from 985 images and 1283 human-generated question-answers. We also thoroughly evaluate QUILT-LLAVA using public histopathology datasets where QUILT-LLAVA significantly outperforms SOTA by over 10% on relative GPT-4 score and 4% and 9% on open and closed set VQA.
****
Traffic Scene Parsing through the TSP6K Dataset
Peng-Tao Jiang, Yuqi Yang, Yang Cao, Qibin Hou, Ming-Ming Cheng, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21874-21885
Traffic scene perception in computer vision is a critically important task to achieve intelligent cities. To date most existing datasets focus on autonomous driving scenes. We observe that the models trained on those driving datasets often yield unsatisfactory results on traffic monitoring scenes. However little effort has been put into improving the traffic monitoring scene understanding mainly due to the lack of specific datasets. To fill this gap we introduce a specialized traffic monitoring dataset termed TSP6K containing images from the traffic monitoring scenario with high-quality pixel-level and instance-level annotations. The TSP6K dataset captures more crowded traffic scenes with several times more traffic participants than the existing driving scenes. We perform a detailed analysis of the dataset and comprehensively evaluate previous popular scene parsing methods instance segmentation methods and unsupervised domain adaption methods. Fu

rthermore considering the vast difference in instance sizes we propose a detail refining decoder for scene parsing which recovers the details of different semantic regions in traffic scenes owing to the proposed TSP6K dataset. Experiments show its effectiveness in parsing the traffic monitoring scenes. Code and dataset are available at https://github.com/PengtaoJiang/TSP6K.
****

Style Aligned Image Generation via Shared Attention
Amir Hertz, Andrey Voynov, Shlomi Fruchter, Daniel Cohen-Or; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4775-4785
Large-scale Text-to-Image (T2I) models have rapidly gained prominence across creative fields generating visually compelling outputs from textual prompts. However controlling these models to ensure consistent style remains challenging with existing methods necessitating fine-tuning and manual intervention to disentangle content and style. In this paper we introduce StyleAligned a novel technique designed to establish style alignment among a series of generated images. By employing minimal `attention sharing' during the diffusion process our method maintains style consistency across images within T2I models. This approach allows for the creation of style-consistent images using a reference style through a straightforward inversion operation. Our method's evaluation across diverse styles and text prompts demonstrates high-quality synthesis and fidelity underscoring its efficacy in achieving consistent style across various inputs.
****

E-GPS: Explainable Geometry Problem Solving via Top-Down Solver and Bottom-Up Generator
Wenjun Wu, Lingling Zhang, Jun Liu, Xi Tang, Yaxian Wang, Shaowei Wang, Qianying Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13828-13837
Geometry Problem Solving has drawn growing attention recently due to its application prospects in intelligent education field. However existing methods are still inadequate to meet the needs of practical application suffering from the following limitations: 1) explainability is not ensured which is essential in real teaching scenarios; 2) the small scale and incomplete annotation of existing datasets make it hard for model to comprehend geometric knowledge. To tackle the above problems we propose a novel method called Explainable Geometry Problem Solving (E-GPS). E-GPS first parses the geometric diagram and problem text into unified formal language representations. Then the answer and explainable reasoning and solving steps are obtained by a Top-Down Problem Solver (TD-PS) which innovatively solves the problem from the target and focuses on what is needed. To alleviate the data issues a Bottom-Up Problem Generator (BU-PG) is devised to augment the data set with various well-annotated constructed geometry problems. It enables us to train an enhanced theorem predictor with a better grasp of theorem knowledge which further improves the efficiency of TD-PS. Extensive experiments demonstrate that E-GPS maintains comparable solving performances with fewer steps and provides outstanding explainability.
****

Back to 3D: Few-Shot 3D Keypoint Detection with Back-Projected 2D Features
Thomas Wimmer, Peter Wonka, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4154-4164
With the immense growth of dataset sizes and computing resources in recent years so-called foundation models have become popular in NLP and vision tasks. In this work we propose to explore foundation models for the task of keypoint detection on 3D shapes. A unique characteristic of keypoint detection is that it requires semantic and geometric awareness while demanding high localization accuracy. To address this problem we propose first to back-project features from large pre-trained 2D vision models onto 3D shapes and employ them for this task. We show that we obtain robust 3D features that contain rich semantic information and analyze multiple candidate features stemming from different 2D foundation models. Second we employ a keypoint candidate optimization module which aims to match the average observed distribution of keypoints on the shape and is guided by the bac

k-projected features. The resulting approach achieves a new state of the art for few-shot keypoint detection on the KeyPointNet dataset almost doubling the performance of the previous best methods.
****
Fourier Priors-Guided Diffusion for Zero-Shot Joint Low-Light Enhancement and Deblurring
Xiaoqian Lv, Shengping Zhang, Chenyang Wang, Yichen Zheng, Bineng Zhong, Chongyi Li, Liqiang Nie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25378-25388
Existing joint low-light enhancement and deblurring methods learn pixel-wise mappings from paired synthetic data which results in limited generalization in real-world scenes. While some studies explore the rich generative prior of pre-trained diffusion models they typically rely on the assumed degradation process and cannot handle unknown real-world degradations well. To address these problems we propose a novel zero-shot framework FourierDiff which embeds Fourier priors into a pre-trained diffusion model to harmoniously handle the joint degradation of luminance and structures. FourierDiff is appealing in its relaxed requirements on paired training data and degradation assumptions. The key zero-shot insight is motivated by image characteristics in the Fourier domain: most luminance information concentrates on amplitudes while structure and content information are closely related to phases. Based on this observation we decompose the sampled results of the reverse diffusion process in the Fourier domain and take advantage of the amplitude of the generative prior to align the enhanced brightness with the distribution of natural images. To yield a sharp and content-consistent enhanced result we further design a spatial-frequency alternating optimization strategy to progressively refine the phase of the input. Extensive experiments demonstrate the superior effectiveness of the proposed method especially in real-world scenes.
****
Neural Markov Random Field for Stereo Matching
Tongfan Guan, Chen Wang, Yun-Hui Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5459-5469
Stereo matching is a core task for many computer vision and robotics applications. Despite their dominance in traditional stereo methods the hand-crafted Markov Random Field (MRF) models lack sufficient modeling accuracy compared to end-to-end deep models. While deep learning representations have greatly improved the unary terms of the MRF models the overall accuracy is still severely limited by the hand-crafted pairwise terms and message passing. To address these issues we propose a neural MRF model where both potential functions and message passing are designed using data-driven neural networks. Our fully data-driven model is built on the foundation of variational inference theory to prevent convergence issues and retain stereo MRF's graph inductive bias. To make the inference tractable and scale well to high-resolution images we also propose a Disparity Proposal Network (DPN) to adaptively prune the search space of disparity. The proposed approach ranks 1^ st on both KITTI 2012 and 2015 leaderboards among all published methods while running faster than 100 ms. This approach significantly outperforms prior global methods e.g. lowering D1 metric by more than 50% on KITTI 2015. In addition our method exhibits strong cross-domain generalization and can recover sharp edges. The codes at https://github.com/aeolusguan/NMRF.
****
Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving
Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14749-14759
In autonomous driving predicting future events in advance and evaluating the foreseeable risks empowers autonomous vehicles to plan their actions enhancing safety and efficiency on the road. To this end we propose Drive-WM the first driving world model compatible with existing end-to-end planning models. Through a joint spatial-temporal modeling facilitated by view factorization our model is the f

irst to generate high-fidelity multiview videos. Building on its powerful genera tion ability we showcase the potential of applying the world model for safe driv ing planning for the first time. Our Drive-WM enables driving into multiple futu res based on distinct driving maneuvers and determines the optimal trajectory ac cording to the image-based rewards. Evaluation on real-world driving datasets ve rifies that our method could generate high-quality consistent and controllable m ultiview videos opening up possibilities for real-world simulations and safe pla nning.
****

OpenESS: Event-based Semantic Scene Understanding with Open Vocabularies
Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R. Cottereau, Wei Tsang Ooi; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15686-15698
Event-based semantic segmentation (ESS) is a fundamental yet challenging task fo r event camera sensing. The difficulties in interpreting and annotating event da ta limit its scalability. While domain adaptation from images to event data can help to mitigate this issue there exist data representational differences that r equire additional effort to resolve. In this work for the first time we synergiz e information from image text and event-data domains and introduce OpenESS to en able scalable ESS in an open-world annotation-efficient manner. We achieve this goal by transferring the semantically rich CLIP knowledge from image-text pairs to event streams. To pursue better cross-modality adaptation we propose a frame- to-event contrastive distillation and a text-to-event semantic consistency regul arization. Experimental results on popular ESS benchmarks showed our approach ou tperforms existing methods. Notably we achieve 53.93% and 43.31% mIoU on DDD17 a nd DSEC-Semantic without using either event or frame labels.
****

Do Vision and Language Encoders Represent the World Similarly?
Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, Noel E. O'Connor; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14334-14343
Aligned text-image encoders such as CLIP have become the de-facto model for visi on-language tasks. Furthermore modality-specific encoders achieve impressive per formances in their respective domains. This raises a central question: does an a lignment exist between uni-modal vision and language encoders since they fundame ntally represent the same physical world? Analyzing the latent spaces structure of vision and language models on image-caption benchmarks using the Centered Ker nel Alignment (CKA) we find that the representation spaces of unaligned and alig ned encoders are semantically similar. In the absence of statistical similarity in aligned encoders like CLIP we show that a possible matching of unaligned enco ders exists without any training. We frame this as a seeded graph-matching probl em exploiting the semantic similarity between graphs and propose two methods - a Fast Quadratic Assignment Problem optimization and a novel localized CKA metric -based matching/retrieval. We demonstrate the effectiveness of this on several d ownstream tasks including cross-lingual cross-domain caption matching and image classification. Code available at github.com/mayug/0-shot-llm-vision.
****

MGMap: Mask-Guided Learning for Online Vectorized HD Map Construction
Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, Jianke Zhu; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14812-14821
Currently high-definition (HD) map construction leans towards a lightweight onli ne generation tendency which aims to preserve timely and reliable road scene inf ormation. However map elements contain strong shape priors. Subtle and sparse an notations make current detection-based frameworks ambiguous in locating relevant feature scopes and cause the loss of detailed structures in prediction. To alle viate these problems we propose MGMap a mask-guided approach that effectively hi ghlights the informative regions and achieves precise map element localization b y introducing the learned masks. Specifically MGMap employs learned masks based

on the enhanced multi-scale BEV features from two perspectives. At the instance level we propose the Mask-activated instance (MAI) decoder which incorporates global instance and structural information into instance queries by the activation of instance masks. At the point level a novel position-guided mask patch refinement (PG-MPR) module is designed to refine point locations from a finer-grained perspective enabling the extraction of point-specific patch information. Compared to the baselines our proposed MGMap achieves a notable improvement of around 10 mAP for different input modalities. Extensive experiments also demonstrate that our approach showcases strong robustness and generalization capabilities. Our code can be found at https://github.com/xiaolul2/MGMap.
****
Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild
Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, Chao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25669-25680
We introduce SUPIR (Scaling-UP Image Restoration) a groundbreaking image restoration method that harnesses generative prior and the power of model scaling up. Leveraging multi-modal techniques and advanced generative prior SUPIR marks a significant advance in intelligent and realistic image restoration. As a pivotal catalyst within SUPIR model scaling dramatically enhances its capabilities and demonstrates new potential for image restoration. We collect a dataset comprising 20 million high-resolution high-quality images for model training each enriched with descriptive text annotations. SUPIR provides the capability to restore images guided by textual prompts broadening its application scope and potential. Moreover we introduce negative-quality prompts to further improve perceptual quality. We also develop a restoration-guided sampling method to suppress the fidelity issue encountered in generative-based restoration. Experiments demonstrate SUPIR's exceptional restoration effects and its novel capacity to manipulate restoration through textual prompts.
****
Q-Instruct: Improving Low-level Visual Abilities for Multi-modality Foundation Models
Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, Weisi Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25490-25500
Multi-modality large language models (MLLMs) as represented by GPT-4V have introduced a paradigm shift for visual perception and understanding tasks that a variety of abilities can be achieved within one foundation model. While current MLLMs demonstrate primary low-level visual abilities from the identification of low-level visual attributes (e.g. clarity brightness) to the evaluation on image quality there's still an imperative to further improve the accuracy of MLLMs to substantially alleviate human burdens. To address this we collect the first dataset consisting of human natural language feedback on low-level vision. Each feedback offers a comprehensive description of an image's low-level visual attributes culminating in an overall quality assessment. The constructed Q-Pathway dataset includes 58K detailed human feedbacks on 18973 multi-sourced images with diverse low-level appearance. To ensure MLLMs can adeptly handle diverse queries we further propose a GPT-participated transformation to convert these feedbacks into a rich set of 200K instruction-response pairs termed Q-Instruct. Experimental results indicate that the Q-Instruct consistently elevates various low-level visual capabilities across multiple base models. We anticipate that our datasets can pave the way for a future that foundation models can assist humans on low-level visual tasks.
****
PoseIRM: Enhance 3D Human Pose Estimation on Unseen Camera Settings via Invariant Risk Minimization
Yanlu Cai, Weizhong Zhang, Yuan Wu, Cheng Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2124-2133

Camera-parameter-free multi-view pose estimation is an emerging technique for 3D
 human pose estimation (HPE). They can infer the camera settings implicitly or e
xplicitly to mitigate the depth uncertainty impact showcasing significant potent
ial in real applications. However due to the limited camera setting diversity in
 the available datasets the inferred camera parameters are always simply hardcod
ed into the model during training and not adaptable to the input in inference ma
king the learned models cannot generalize well under unseen camera settings. A n
atural solution is to artificially synthesize some samples i.e. 2D-3D pose pairs
 under massive new camera settings. Unfortunately to prevent over-fitting the ex
isting camera setting the number of synthesized samples for each new camera sett
ing should be comparable with that for the existing one which multiplies the sca
le of training and even makes it computationally prohibitive. In this paper we p
ropose a novel HPE approach under the invariant risk minimization (IRM) paradigm
. Precisely we first synthesize 2D poses from myriad camera settings. We then tr
ain our model under the IRM paradigm which targets at learning a common optimal
model across all camera settings and thus enforces the model to automatically le
arn the camera parameters based on the input data. This allows the model to accu
rately infer 3D poses on unseen data by training on only a handful of samples fr
om each synthesized setting and thus avoid the unbearable training cost incremen
t. Another appealing feature of our method is that benefited from the capability
 of IRM in identifying the invariant features its performance on the seen camera
 settings is enhanced as well. Comprehensive experiments verify the superiority
of our approach.
****
Zero-Shot Structure-Preserving Diffusion Model for High Dynamic Range Tone Mappi
ng
Ruoxi Zhu, Shusong Xu, Peiye Liu, Sicheng Li, Yanheng Lu, Dimin Niu, Zihao Liu,
Zihao Meng, Zhiyong Li, Xinhua Chen, Yibo Fan; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26130-26139
Tone mapping techniques aiming to convert high dynamic range (HDR) images to hig
h-quality low dynamic range (LDR) images for display play a more crucial role in
 real-world vision systems with the increasing application of HDR images. Howeve
r obtaining paired HDR and high-quality LDR images is difficult posing a challen
ge to deep learning based tone mapping methods. To overcome this challenge we pr
opose a novel zero-shot tone mapping framework that utilizes shared structure kn
owledge allowing us to transfer a pre-trained mapping model from the LDR domain
to HDR fields without paired training data. Our approach involves decomposing bo
th the LDR and HDR images into two components: structural information and tonal
information. To preserve the original image's structure we modify the reverse sa
mpling process of a diffusion model and explicitly incorporate the structure inf
ormation into the intermediate results. Additionally for improved image details
we introduce a dual-control network architecture that enables different types of
 conditional inputs to control different scales of the output. Experimental resu
lts demonstrate the effectiveness of our approach surpassing previous state-of-t
he-art methods both qualitatively and quantitatively. Moreover our model exhibit
s versatility and can be applied to other low-level vision tasks without retrain
ing. The code is available at https://github.com/ZSDM-HDR/Zero-Shot-Diffusion-HD
R.
****
VidLA: Video-Language Alignment at Scale
Mamshad Nayeem Rizve, Fan Fei, Jayakrishnan Unnikrishnan, Son Tran, Benjamin Z.
Yao, Belinda Zeng, Mubarak Shah, Trishul Chilimbi; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14043-140
55
In this paper we propose VidLA an approach for video-language alignment at scale
. There are two major limitations of previous video-language alignment approache
s. First they do not capture both short-range and long-range temporal dependenci
es and typically employ complex hierarchical deep network architectures that are
 hard to integrate with existing pretrained image-text foundation models. To eff
ectively address this limitation we instead keep the network architecture simple

and use a set of data tokens that operate at different temporal resolutions in a hierarchical manner accounting for the temporally hierarchical nature of videos. By employing a simple two-tower architecture we are able to initialize our video-language model with pretrained image-text foundation models thereby boosting the final performance. Second existing video-language alignment works struggle due to the lack of semantically aligned large-scale training data. To overcome it we leverage recent LLMs to curate the largest video-language dataset to date with better visual grounding. Furthermore unlike existing video-text datasets which only contain short clips our dataset is enriched with video clips of varying durations to aid our temporally hierarchical data tokens in extracting better representations at varying temporal scales. Overall empirical results show that our proposed approach surpasses state-of-the-art methods on multiple retrieval benchmarks especially on longer videos and performs competitively on classification benchmarks.
****
VoCo: A Simple-yet-Effective Volume Contrastive Learning Framework for 3D Medical Image Analysis
Linshan Wu, Jiaxin Zhuang, Hao Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22873-22882
Self-Supervised Learning (SSL) has demonstrated promising results in 3D medical image analysis. However the lack of high-level semantics in pre-training still heavily hinders the performance of downstream tasks. We observe that 3D medical images contain relatively consistent contextual position information i.e. consistent geometric relations between different organs which leads to a potential way for us to learn consistent semantic representations in pre-training. In this paper we propose a simple-yet-effective Volume Contrast (VoCo) framework to leverage the contextual position priors for pre-training. Specifically we first generate a group of base crops from different regions while enforcing feature discrepancy among them where we employ them as class assignments of different regions. Then we randomly crop sub-volumes and predict them belonging to which class (located at which region) by contrasting their similarity to different base crops which can be seen as predicting contextual positions of different sub-volumes. Through this pretext task VoCo implicitly encodes the contextual position priors into model representations without the guidance of annotations enabling us to effectively improve the performance of downstream tasks that require high-level semantics. Extensive experimental results on six downstream tasks demonstrate the superior effectiveness of VoCo. Code will be available at https://github.com/Luffy03/VoCo.
****
CCEdit: Creative and Controllable Video Editing via Diffusion Models
Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6712-6722
In this paper we present CCEdit a versatile generative video editing framework based on diffusion models. Our approach employs a novel trident network structure that separates structure and appearance control ensuring precise and creative editing capabilities. Utilizing the foundational ControlNet architecture we maintain the structural integrity of the video during editing. The incorporation of an additional appearance branch enables users to exert fine-grained control over the edited key frame. These two side branches seamlessly integrate into the main branch which is constructed upon existing text-to-image (T2I) generation models through learnable temporal layers. The versatility of our framework is demonstrated through a diverse range of choices in both structure representations and personalized T2I models as well as the option to provide the edited key frame. To facilitate comprehensive evaluation we introduce the BalanceCC benchmark dataset comprising 100 videos and 4 target prompts for each video. Our extensive user studies compare CCEdit with eight state-of-the-art video editing methods. The outcomes demonstrate CCEdit's substantial superiority over all other methods.
****
IPoD: Implicit Field Learning with Point Diffusion for Generalizable 3D Object R

econstruction from Single RGB-D Images
Yushuang Wu, Luyue Shi, Junhao Cai, Weihao Yuan, Lingteng Qiu, Zilong Dong, Lief eng Bo, Shuguang Cui, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20432-20442
Generalizable 3D object reconstruction from single-view RGB-D images remains a challenging task particularly with real-world data. Current state-of-the-art methods develop Transformer-based implicit field learning necessitating an intensive learning paradigm that requires dense query-supervision uniformly sampled throughout the entire space. We propose a novel approach IPoD which harmonizes implicit field learning with point diffusion. This approach treats the query points for implicit field learning as a noisy point cloud for iterative denoising allowing for their dynamic adaptation to the target object shape. Such adaptive query points harness diffusion learning's capability for coarse shape recovery and also enhances the implicit representation's ability to delineate finer details. Besides an additional self-conditioning mechanism is designed to use implicit predictions as the guidance of diffusion learning leading to a cooperative system. Experiments conducted on the CO3D-v2 dataset affirm the superiority of IPoD achieving 7.8% improvement in F-score and 28.6% in Chamfer distance over existing methods. The generalizability of IPoD is also demonstrated on the MVImgNet dataset. Our project page is at https://yushuang-wu.github.io/IPoD.
****
HAVE-FUN: Human Avatar Reconstruction from Few-Shot Unconstrained Images
Xihe Yang, Xingyu Chen, Daiheng Gao, Shaohui Wang, Xiaoguang Han, Baoyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 742-752
As for human avatar reconstruction contemporary techniques commonly necessitate the acquisition of costly data and struggle to achieve satisfactory results from a small number of casual images. In this paper we investigate this task from a few-shot unconstrained photo album. The reconstruction of human avatars from such data sources is challenging because of limited data amount and dynamic articulated poses. For handling dynamic data we integrate a skinning mechanism with deep marching tetrahedra (DMTet) to form a drivable tetrahedral representation which drives arbitrary mesh topologies generated by the DMTet for the adaptation of unconstrained images. To effectively mine instructive information from few-shot data we devise a two-phase optimization method with few-shot reference and few-shot guidance. The former focuses on aligning avatar identity with reference images while the latter aims to generate plausible appearances for unseen regions. Overall our framework called HaveFun can undertake avatar reconstruction rendering and animation. Extensive experiments on our developed benchmarks demonstrate that HaveFun exhibits substantially superior performance in reconstructing the human body and hand.
****
ERMVP: Communication-Efficient and Collaboration-Robust Multi-Vehicle Perception in Challenging Environments
Jingyu Zhang, Kun Yang, Yilei Wang, Hanqi Wang, Peng Sun, Liang Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12575-12584
Collaborative perception enhances perception performance by enabling autonomous vehicles to exchange complementary information. Despite its potential to revolutionize the mobile industry challenges in various environments such as communication bandwidth limitations localization errors and information aggregation inefficiencies hinder its implementation in practical applications. In this work we propose ERMVP a communication-Efficient and collaboration-Robust Multi-Vehicle Perception method in challenging environments. Specifically ERMVP has three distinct strengths: i) It utilizes the hierarchical feature sampling strategy to abstract a representative set of feature vectors using less communication overhead for efficient communication; ii) It employs the sparse consensus features to execute precise spatial location calibrations effectively mitigating the implications of vehicle localization errors; iii) A pioneering feature fusion and interaction paradigm is introduced to integrate holistic spatial semantics among different

vehicles and data sources. To thoroughly validate our method we conduct extensiv e experiments on real-world and simulated datasets. The results demonstrate that the proposed ERMVP is significantly superior to the state-of-the-art collaborat ive perception methods.
****
DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing
Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, Xingang Pan; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 79 12-7921
Diffusion models have achieved remarkable image generation quality surpassing pr evious generative models. However a notable limitation of diffusion models in co mparison to GANs is their difficulty in smoothly interpolating between two image samples due to their highly unstructured latent space. Such a smooth interpolat ion is intriguing as it naturally serves as a solution for the image morphing ta sk with many applications. In this work we address this limitation via DiffMorph er an approach that enables smooth and natural image interpolation by harnessing the prior knowledge of a pre-trained diffusion model. Our key idea is to captur e the semantics of the two images by fitting two LoRAs to them respectively and interpolate between both the LoRA parameters and the latent noises to ensure a s mooth semantic transition where correspondence automatically emerges without the need for annotation. In addition we propose an attention interpolation and inje ction technique an adaptive normalization adjustment method and a new sampling s chedule to further enhance the smoothness between consecutive images. Extensive experiments demonstrate that DiffMorpher achieves starkly better image morphing effects than previous methods across a variety of object categories bridging a c ritical functional gap that distinguished diffusion models from GANs.
****
Towards Real-World HDR Video Reconstruction: A Large-Scale Benchmark Dataset and A Two-Stage Alignment Network
Yong Shu, Liquan Shen, Xiangyu Hu, Mengyao Li, Zihao Zhou; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2 879-2888
As an important and practical way to obtain high dynamic range (HDR) video HDR v ideo reconstruction from sequences with alternating exposures is still less expl ored mainly due to the lack of large-scale real-world datasets. Existing methods are mostly trained on synthetic datasets which perform poorly in real scenes. I n this work to facilitate the development of real-world HDR video reconstruction we present Real-HDRV a large-scale real-world benchmark dataset for HDR video r econstruction featuring various scenes diverse motion patterns and high-quality labels. Specifically our dataset contains 500 LDRs-HDRs video pairs comprising a bout 28000 LDR frames and 4000 HDR labels covering daytime nighttime indoor and outdoor scenes. To our best knowledge our dataset is the largest real-world HDR video reconstruction dataset. Correspondingly we propose an end-to-end network f or HDR video reconstruction where a novel two-stage strategy is designed to perf orm alignment sequentially. Specifically the first stage performs global alignme nt with the adaptively estimated global offsets reducing the difficulty of subse quent alignment. The second stage implicitly performs local alignment in a coars e-to-fine manner at the feature level using the adaptive separable convolution. Extensive experiments demonstrate that: (1) models trained on our dataset can ac hieve better performance on real scenes than those trained on synthetic datasets ; (2) our method outperforms previous state-of-the-art methods. Our dataset is a vailable at https://github.com/yungsyu99/Real-HDRV.
****
Efficient 3D Implicit Head Avatar with Mesh-anchored Hash Table Blendshapes
Ziqian Bai, Feitong Tan, Sean Fanello, Rohit Pandey, Mingsong Dou, Shichen Liu, Ping Tan, Yinda Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1975-1984
3D head avatars built with neural implicit volumetric representations have achie ved unprecedented levels of photorealism. However the computational cost of thes e methods remains a significant barrier to their widespread adoption particularl

y in real-time applications such as virtual reality and teleconferencing. While attempts have been made to develop fast neural rendering approaches for static scenes these methods cannot be simply employed to support realistic facial expressions such as in the case of a dynamic facial performance. To address these challenges we propose a novel fast 3D neural implicit head avatar model that achieves real-time rendering while maintaining fine-grained controllability and high rendering quality. Our key idea lies in the introduction of local hash table blendshapes which are learned and attached to the vertices of an underlying face parametric model. These per-vertex hash-tables are linearly merged with weights predicted via a CNN resulting in expression dependent embeddings. Our novel representation enables efficient density and color predictions using a lightweight MLP which is further accelerated by a hierarchical nearest neighbor search method. Extensive experiments show that our approach runs in real-time while achieving comparable rendering quality to state-of-the-arts and decent results on challenging expressions.
****
PikeLPN: Mitigating Overlooked Inefficiencies of Low-Precision Neural Networks
Marina Neseem, Conor McCullough, Randy Hsin, Chas Leichner, Shan Li, In Suk Chong, Andrew Howard, Lukasz Lew, Sherief Reda, Ville-Mikko Rautio, Daniele Moro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15996-16005
Low-precision quantization is recognized for its efficacy in neural network optimization. Our analysis reveals that non-quantized elementwise operations which are prevalent in layers such as parameterized activation functions batch normalization and quantization scaling dominate the inference cost of low-precision models. These non-quantized elementwise operations are commonly overlooked in SOTA efficiency metrics such as Arithmetic Computation Effort (ACE). In this paper we propose ACEv2 - an extended version of ACE which offers a better alignment with the inference cost of quantized models and their energy consumption on ML hardware. Moreover we introduce PikeLPN a model that addresses these efficiency issues by applying quantization to both elementwise operations and multiply-accumulate operations. In particular we present a novel quantization technique for batch normalization layers named QuantNorm which allows for quantizing the batch normalization parameters without compromising the model performance. Additionally we propose applying Double Quantization where the quantization scaling parameters are quantized. Furthermore we recognize and resolve the issue of distribution mismatch in Separable Convolution layers by introducing Distribution-Heterogeneous Quantization which enables quantizing them to low-precision. PikeLPN achieves Pareto-optimality in efficiency-accuracy trade-off with up to 3X efficiency improvement compared to SOTA low-precision models.
****
CurveCloudNet: Processing Point Clouds with 1D Structure
Colton Stearns, Alex Fu, Jiateng Liu, Jeong Joon Park, Davis Rempe, Despoina Paschalidou, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27981-27991
Modern depth sensors such as LiDAR operate by sweeping laser-beams across the scene resulting in a point cloud with notable 1D curve-like structures. In this work we introduce a new point cloud processing scheme and backbone called CurveCloudNet which takes advantage of the curve-like structure inherent to these sensors. While existing backbones discard the rich 1D traversal patterns and rely on generic 3D operations CurveCloudNet parameterizes the point cloud as a collection of polylines (dubbed a "curve cloud") establishing a local surface-aware ordering on the points. By reasoning along curves CurveCloudNet captures lightweight curve-aware priors to efficiently and accurately reason in several diverse 3D environments. We evaluate CurveCloudNet on multiple synthetic and real datasets that exhibit distinct 3D size and structure. We demonstrate that CurveCloudNet outperforms both point-based and sparse-voxel backbones in various segmentation settings notably scaling to large scenes better than point-based alternatives while exhibiting improved single-object performance over sparse-voxel alternatives. In all CurveCloudNet is an efficient and accurate backbone that can handle a large

r variety of 3D environments than past works.
****
CAGE: Controllable Articulation GEneration
Jiayi Liu, Hou In Ivan Tam, Ali Mahdavi-Amiri, Manolis Savva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17880-17889
We address the challenge of generating 3D articulated objects in a controllable fashion. Currently modeling articulated 3D objects is either achieved through laborious manual authoring or using methods from prior work that are hard to scale and control directly. We leverage the interplay between part shape connectivity and motion using a denoising diffusion-based method with attention modules designed to extract correlations between part attributes. Our method takes an object category label and a part connectivity graph as input and generates an object's geometry and motion parameters. The generated objects conform to user-specified constraints on the object category part shape and part articulation. Our experiments show that our method outperforms the state-of-the-art in articulated object generation producing more realistic objects while conforming better to user constraints.
****
No Time to Train: Empowering Non-Parametric Networks for Few-shot 3D Scene Segmentation
Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Jiaming Liu, Han Xiao, Chaoyou Fu, Hao Dong, Peng Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3838-3847
To reduce the reliance on large-scale datasets recent works in 3D segmentation resort to few-shot learning. Current 3D few-shot segmentation methods first pre-train models on 'seen' classes and then evaluate their generalization performance on 'unseen' classes. However the prior pre-training stage not only introduces excessive time overhead but also incurs a significant domain gap on 'unseen' classes. To tackle these issues we propose a Non-parametric Network for few-shot 3D Segmentation Seg-NN and its Parametric variant Seg-PN. Without training Seg-NN extracts dense representations by hand-crafted filters and achieves comparable performance to existing parameterized models. Due to the elimination of pre-training Seg-NN can alleviate the domain gap issue and save a substantial amount of time. Based on Seg-NN Seg-PN only requires training a lightweight QUEry-Support Transferring (QUEST) module which enhances the interaction between the support set and query set. Experiments suggest that Seg-PN outperforms previous state-of-the-art method by +4.19% and +7.71% mIoU on S3DIS and ScanNet datasets respectively while reducing training time by -90% indicating its effectiveness and efficiency. Code is available https://github.com/yangyangyang127/Seg-NN.
****
PhysGaussian: Physics-Integrated 3D Gaussians for Generative Dynamics
Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, Chenfanfu Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4389-4398
We introduce PhysGaussian a new method that seamlessly integrates physically grounded Newtonian dynamics within 3D Gaussians to achieve high-quality novel motion synthesis. Employing a customized Material Point Method (MPM) our approach enriches 3D Gaussian kernels with physically meaningful kinematic deformation and mechanical stress attributes all evolved in line with continuum mechanics principles. A defining characteristic of our method is the seamless integration between physical simulation and visual rendering: both components utilize the same 3D Gaussian kernels as their discrete representations. This negates the necessity for triangle/tetrahedron meshing marching cubes cage meshes or any other geometry embedding highlighting the principle of "what you see is what you simulate (WS^2)". Our method demonstrates exceptional versatility across a wide variety of materials--including elastic entities plastic metals non-Newtonian fluids and granular materials--showcasing its strong capabilities in creating diverse visual content with novel viewpoints and movements.
****

Spatio-Temporal Turbulence Mitigation: A Translational Perspective
Xingguang Zhang, Nicholas Chimitt, Yiheng Chi, Zhiyuan Mao, Stanley H. Chan; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2024, pp. 2889-2899
Recovering images distorted by atmospheric turbulence is a challenging inverse p
roblem due to the stochastic nature of turbulence. Although numerous turbulence
mitigation (TM) algorithms have been proposed their efficiency and generalizatio
n to real-world dynamic scenarios remain severely limited. Building upon the int
uitions of classical TM algorithms we present the Deep Atmospheric TUrbulence Mi
tigation network (DATUM). DATUM aims to overcome major challenges when transitio
ning from classical to deep learning approaches. By carefully integrating the me
rits of classical multi-frame TM methods into a deep network structure we demons
trate that DATUM can efficiently perform long-range temporal aggregation using a
 recurrent fashion while deformable attention and temporal-channel attention sea
mlessly facilitate pixel registration and lucky imaging. With additional supervi
sion tilt and blur degradation can be jointly mitigated. These inductive biases
empower DATUM to significantly outperform existing methods while delivering a te
nfold increase in processing speed. A large-scale training dataset ATSyn is pres
ented as a co-invention to enable the generalization to real turbulence. Our cod
e and datasets are available at https://xg416.github.io/DATUM/
****
FocusMAE: Gallbladder Cancer Detection from Ultrasound Videos with Focused Maske
d Autoencoders
Soumen Basu, Mayuna Gupta, Chetan Madan, Pankaj Gupta, Chetan Arora; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 11715-11725
In recent years automated Gallbladder Cancer (GBC) detection has gained the atte
ntion of researchers. Current state-of-the-art (SOTA) methodologies relying on u
ltrasound sonography (US) images exhibit limited generalization emphasizing the
need for transformative approaches. We observe that individual US frames may lac
k sufficient information to capture disease manifestation. This study advocates
for a paradigm shift towards video-based GBC detection leveraging the inherent a
dvantages of spatiotemporal representations. Employing the Masked Autoencoder (M
AE) for representation learning we address shortcomings in conventional image-ba
sed methods. We propose a novel design called FocusMAE to systematically bias th
e selection of masking tokens from high-information regions fostering a more ref
ined representation of malignancy. Additionally we contribute the most extensive
 US video dataset for GBC detection. We also note that this is the first study o
n US video-based GBC detection. We validate the proposed methods on the curated
dataset and report a new SOTA accuracy of 96.4% for the GBC detection problem ag
ainst an accuracy of 84% by current Image-based SOTA - GBCNet and RadFormer and
94.7% by Video-based SOTA - AdaMAE. We further demonstrate the generality of the
 proposed FocusMAE on a public CT-based Covid detection dataset reporting an imp
rovement in accuracy by 3.3% over current baselines. Project page with source co
de trained models and data is available at: https://gbc-iitd.github.io/focusmae.
****
Grounded Text-to-Image Synthesis with Attention Refocusing
Quynh Phung, Songwei Ge, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7932-7942
Driven by the scalable diffusion models trained on large-scale datasets text-to-
image synthesis methods have shown compelling results. However these models stil
l fail to precisely follow the text prompt involving multiple objects attributes
 or spatial compositions. In this paper we reveal the potential causes of the di
ffusion model's cross-attention and self-attention layers. We propose two novel
losses to refocus attention maps according to a given spatial layout during samp
ling. Creating the layouts manually requires additional effort and can be tediou
s. Therefore we explore using large language models (LLM) to produce these layou
ts for our method. We conduct extensive experiments on the DrawBench HRS and TIF
A benchmarks to evaluate our proposed method. We show that our proposed attentio
n refocusing effectively improves the controllability of existing approaches.

****
OpenStreetView-5M: The Many Roads to Global Visual Geolocation
Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, Loic Landrieu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21967-21977
Determining the location of an image anywhere on Earth is a complex visual task which makes it particularly relevant for evaluating computer vision algorithms. Determining the location of an image anywhere on Earth is a complex visual task which makes it particularly relevant for evaluating computer vision algorithms. Yet the absence of standard large-scale open-access datasets with reliably localizable images has limited its potential. To address this issue we introduce OpenStreetView-5M a large-scale open-access dataset comprising over 5.1 million geo-referenced street view images covering 225 countries and territories. In contrast to existing benchmarks we enforce a strict train/test separation allowing us to evaluate the relevance of learned geographical features beyond mere memorization. To demonstrate the utility of our dataset we conduct an extensive benchmark of various state-of-the-art image encoders spatial representations and training strategies. All associated codes and models can be found at https://github.com/gastruc/osv5m.
****
Visual Concept Connectome (VCC): Open World Concept Discovery and their Interlayer Connections in Deep Models
Matthew Kowal, Richard P. Wildes, Konstantinos G. Derpanis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10895-10905
Understanding what deep network models capture in their learned representations is a fundamental challenge in computer vision. We present a new methodology to understanding such vision models the Visual Concept Connectome (VCC) which discovers human interpretable concepts and their interlayer connections in a fully unsupervised manner. Our approach simultaneously reveals fine-grained concepts at a layer connection weightings across all layers and is amendable to global analysis of network structure (e.g. branching pattern of hierarchical concept assemblies). Previous work yielded ways to extract interpretable concepts from single layers and examine their impact on classification but did not afford multilayer concept analysis across an entire network architecture. Quantitative and qualitative empirical results show the effectiveness of VCCs in the domain of image classification. Also we leverage VCCs for the application of failure mode debugging to reveal where mistakes arise in deep networks.
****
IReNe: Instant Recoloring of Neural Radiance Fields
Alessio Mazzucchelli, Adrian Garcia-Garcia, Elena Garces, Fernando Rivas-Manzaneque, Francesc Moreno-Noguer, Adrian Penate-Sanchez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5937-5946
Advances in NERFs have allowed for 3D scene reconstructions and novel view synthesis. Yet efficiently editing these representations while retaining photorealism is an emerging challenge. Recent methods face three primary limitations: they're slow for interactive use lack precision at object boundaries and struggle to ensure multi-view consistency. We introduce IReNe to address these limitations enabling swift near real-time color editing in NeRF. Leveraging a pre-trained NeRF model and a single training image with user-applied color edits IReNe swiftly adjusts network parameters in seconds. This adjustment allows the model to generate new scene views accurately representing the color changes from the training image while also controlling object boundaries and view-specific effects. Object boundary control is achieved by integrating a trainable segmentation module into the model. The process gains efficiency by retraining only the weights of the last network layer. We observed that neurons in this layer can be classified into those responsible for view-dependent appearance and those contributing to diffuse appearance. We introduce an automated classification approach to identify the

se neuron types and exclusively fine-tune the weights of the diffuse neurons. Th
is further accelerates training and ensures consistent color edits across differ
ent views. A thorough validation on a new dataset with edited object colors show
s significant quantitative and qualitative advancements over competitors acceler
ating speeds by 5x and 500x.
****
Class Tokens Infusion for Weakly Supervised Semantic Segmentation
Sung-Hoon Yoon, Hoyong Kwon, Hyeonseong Kim, Kuk-Jin Yoon; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3
595-3605
Weakly Supervised Semantic Segmentation (WSSS) relies on Class Activation Maps (
CAMs) to extract spatial information from image-level labels. With the success o
f Vision Transformer (ViT) the migration of ViT is actively conducted in WSSS. T
his work proposes a novel WSSS framework with Class Token Infusion (CTI). By inf
using the class tokens from images we guide class tokens to possess class-specif
ic distinct characteristics and global-local consistency. For this we devise two
 kinds of token infusion: 1) Intra-image Class Token Infusion (I-CTI) and 2) Cro
ss-Image Class Token Infusion (C-CTI). In I-CTI we infuse the class tokens from
the same but differently augmented images and thus make CAMs consistent among va
rious deformations (view color). In C-CTI by infusing the class tokens from the
other images and imposing the resulting CAMs to be similar it learns class-speci
fic distinct characteristics. Besides the CTI we bring the background (BG) conce
pt into ViT with the BG token to reduce the false positive activation of CAMs. W
e demonstrate the effectiveness of our method on PASCAL VOC 2012 and MS COCO 201
4 datasets achieving state-of-the-art results in weakly supervised semantic segm
entation. The code is available at https://github.com/yoon307/CTI
****
FedHCA2: Towards Hetero-Client Federated Multi-Task Learning
Yuxiang Lu, Suizhi Huang, Yuwen Yang, Shalayiding Sirejiding, Yue Ding, Hongtao
Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2024, pp. 5599-5609
Federated Learning (FL) enables joint training across distributed clients using
their local data privately. Federated Multi-Task Learning (FMTL) builds on FL to
 handle multiple tasks assuming model congruity that identical model architectur
e is deployed in each client. To relax this assumption and thus extend real-worl
d applicability we introduce a novel problem setting Hetero-Client Federated Mul
ti-Task Learning (HC-FMTL) to accommodate diverse task setups. The main challeng
e of HC-FMTL is the model incongruity issue that invalidates conventional aggreg
ation methods. It also escalates the difficulties in model aggregation to deal w
ith data and task heterogeneity inherent in FMTL. To address these challenges we
 propose the FedHCA^2 framework which allows for federated training of personali
zed models by modeling relationships among heterogeneous clients. Drawing on our
 theoretical insights into the difference between multi-task and federated optim
ization we propose the Hyper Conflict-Averse Aggregation scheme to mitigate conf
licts during encoder updates. Additionally inspired by task interaction in MTL t
he Hyper Cross Attention Aggregation scheme uses layer-wise cross attention to e
nhance decoder interactions while alleviating model incongruity. Moreover we emp
loy learnable Hyper Aggregation Weights for each client to customize personalize
d parameter updates. Extensive experiments demonstrate the superior performance
of FedHCA^2 in various HC-FMTL scenarios compared to representative methods. Cod
e is available at https://github.com/innovator-zero/FedHCA2.
****
Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive
 Image Fusion
Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, Jiayi Ma; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27026
-27035
Image fusion aims to combine information from different source images to create
a comprehensively representative image. Existing fusion methods are typically he
lpless in dealing with degradations in low-quality source images and non-interac

tive to multiple subjective and objective needs. To solve them we introduce a no vel approach that leverages semantic text guidance image fusion model for degrad ation-aware and interactive image fusion task termed as Text-IF. It innovatively extends the classical image fusion to the text guided image fusion along with t he ability to harmoniously address the degradation and interaction issues during fusion. Through the text semantic encoder and semantic interaction fusion decod er Text-IF is accessible to the all-in-one infrared and visible image degradatio n-aware processing and the interactive flexible fusion outcomes. In this way Tex t-IF achieves not only multi-modal image fusion but also multi-modal information fusion. Extensive experiments prove that our proposed text guided image fusion strategy has obvious advantages over SOTA methods in the image fusion performanc e and degradation treatment. The code is available at https://github.com/Xunpeng Yi/Text-IF.
****
GRAM: Global Reasoning for Multi-Page VQA
Tsachi Blau, Sharon Fogel, Roi Ronen, Alona Golts, Roy Ganz, Elad Ben Avraham, A viad Aberdam, Shahar Tsiper, Ron Litman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15598-15607
The increasing use of transformer-based large language models brings forward the challenge of processing long sequences. In document visual question answering (DocVQA) leading methods focus on the single-page setting while documents can spa n hundreds of pages. We present GRAM a method that seamlessly extends pre-traine d single-page models to the multi-page setting without requiring computationally -heavy pretraining. To do so we leverage a single-page encoder for local page-le vel understanding and enhance it with document-level designated layers and learn able tokens facilitating the flow of information across pages for global reasoni ng. To enforce our model to utilize the newly introduced document tokens we prop ose a tailored bias adaptation method. For additional computational savings duri ng decoding we introduce an optional compression stage using our compression-tra nsformer (CFormer)reducing the encoded sequence length thereby allowing a tradeo ff between quality and latency. Extensive experiments showcase GRAM's state-of-t he-art performance on the benchmarks for multi-page DocVQA demonstrating the eff ectiveness of our approach.
****
MS-DETR: Efficient DETR Training with Mixed Supervision
Chuyang Zhao, Yifan Sun, Wenhao Wang, Qiang Chen, Errui Ding, Yi Yang, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2024, pp. 17027-17036
DETR accomplishes end-to-end object detection through iteratively generating mul tiple object candidates based on image features and promoting one candidate for each ground-truth object. The traditional training procedure using one-to-one su pervision in the original DETR lacks direct supervision for the object detection candidates. We aim at improving the DETR training efficiency by explicitly supe rvising the candidate generation procedure through mixing one-to-one supervision and one-to-many supervision. Our approach namely MS-DETR is simple and places o ne-to-many supervision to the object queries of the primary decoder that is used for inference. In comparison to existing DETR variants with one-to-many supervi sion such as Group DETR and Hybrid DETR our approach does not need additional de coder branches or object queries. The object queries of the primary decoder in o ur approach directly benefit from one-to-many supervision and thus are superior in object candidate prediction. Experimental results show that our approach outp erforms related DETR variants such as DN-DETR Hybrid DETR and Group DETR and the combination with related DETR variants further improves the performance.
****
Learning to Produce Semi-dense Correspondences for Visual Localization
Khang Truong Giang, Soohwan Song, Sungho Jo; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19468-19478
This study addresses the challenge of performing visual localization in demandin g conditions such as night-time scenarios adverse weather and seasonal changes. While many prior studies have focused on improving image matching performance to

 facilitate reliable dense keypoint matching between images existing methods oft
en heavily rely on predefined feature points on a reconstructed 3D model. Conseq
uently they tend to overlook unobserved keypoints during the matching process. T
herefore dense keypoint matches are not fully exploited leading to a notable red
uction in accuracy particularly in noisy scenes. To tackle this issue we propose
 a novel localization method that extracts reliable semi-dense 2D-3D matching po
ints based on dense keypoint matches. This approach involves regressing semi-den
se 2D keypoints into 3D scene coordinates using a point inference network. The n
etwork utilizes both geometric and visual cues to effectively infer 3D coordinat
es for unobserved keypoints from the observed ones. The abundance of matching in
formation significantly enhances the accuracy of camera pose estimation even in
scenarios involving noisy or sparse 3D models. Comprehensive evaluations demonst
rate that the proposed method outperforms other methods in challenging scenes an
d achieves competitive results in large-scale visual localization benchmarks. Th
e code will be available at https://github.com/TruongKhang/DeViLoc
****
Amodal Ground Truth and Completion in the Wild
Guanqi Zhan, Chuanxia Zheng, Weidi Xie, Andrew Zisserman; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28
003-28013
This paper studies amodal image segmentation: predicting entire object segmentat
ion masks including both visible and invisible (occluded) parts. In previous wor
k the amodal segmentation ground truth on real images is usually predicted by ma
nual annotaton and thus is subjective. In contrast we use 3D data to establish a
n automatic pipeline to determine authentic ground truth amodal masks for partia
lly occluded objects in real images. This pipeline is used to construct an amoda
l completion evaluation benchmark MP3D-Amodal consisting of a variety of object
categories and labels. To better handle the amodal completion task in the wild w
e explore two architecture variants: a two-stage model that first infers the occ
luder followed by amodal mask completion; and a one-stage model that exploits th
e representation power of Stable Diffusion for amodal segmentation across many c
ategories. Without bells and whistles our method achieves a new state-of-the-art
 performance on Amodal segmentation datasets that cover a large variety of objec
ts including COCOA and our new MP3D-Amodal dataset. The dataset model and code a
re available at https://www. robots.ox.ac.uk/ vgg/research/amodal/.
****
Motion Diversification Networks
Hee Jae Kim, Eshed Ohn-Bar; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2024, pp. 1650-1660
We introduce Motion Diversification Networks a novel framework for learning to g
enerate realistic and diverse 3D human motion. Despite recent advances in deep g
enerative motion modeling existing models often fail to produce samples that cap
ture the full range of plausible and natural 3D human motion within a given cont
ext. The lack of diversity becomes even more apparent in applications where subt
le and multi-modal 3D human forecasting is crucial for safety such as robotics a
nd autonomous driving. Towards more realistic and functional 3D motion models we
 highlight limitations in existing generative modeling techniques particularly i
n overly simplistic latent code sampling strategies. We then introduce a transfo
rmer-based diversification mechanism that learns to effectively guide sampling i
n the latent space. Our proposed attention-based module queries multiple stochas
tic samples to flexibly predict a diverse set of latent codes which can be subse
quently decoded into motion samples. The proposed framework achieves state-of-th
e-art diversity and accuracy prediction performance across a range of benchmarks
 and settings particularly when used to forecast intricate in-the-wild 3D human
motion within complex urban environments. Our models datasets and code are avail
able at https://mdncvpr.github.io/.
****
Telling Left from Right: Identifying Geometry-Aware Semantic Correspondence
Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun,
 Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR), 2024, pp. 3076-3085
While pre-trained large-scale vision models have shown significant promise for semantic correspondence their features often struggle to grasp the geometry and orientation of instances. This paper identifies the importance of being geometry-aware for semantic correspondence and reveals a limitation of the features of current foundation models under simple post-processing. We show that incorporating this information can markedly enhance semantic correspondence performance with simple but effective solutions in both zero-shot and supervised settings. We also construct a new challenging benchmark for semantic correspondence built from an existing animal pose estimation dataset for both pre-training validating models. Our method achieves a PCK@0.10 score of 65.4 (zero-shot) and 85.6 (supervised) on the challenging SPair-71k dataset outperforming the state of the art by 5.5p and 11.0p absolute gains respectively. Our code and datasets are publicly available at: https://telling-left-from-right.github.io.
****
NECA: Neural Customizable Human Avatar
Junjin Xiao, Qing Zhang, Zhan Xu, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20091-20101
Human avatar has become a novel type of 3D asset with various applications. Ideally a human avatar should be fully customizable to accommodate different settings and environments. In this work we introduce NECA an approach capable of learning versatile human representation from monocular or sparse-view videos enabling granular customization across aspects such as pose shadow shape lighting and texture. The core of our approach is to represent humans in complementary dual spaces and predict disentangled neural fields of geometry albedo shadow as well as an external lighting from which we are able to derive realistic rendering with high-frequency details via volumetric rendering. Extensive experiments demonstrate the advantage of our method over the state-of-the-art methods in photorealistic rendering as well as various editing tasks such as novel pose synthesis and relighting. Our code is available at https://github.com/iSEE-Laboratory/NECA.
****
BEVSpread: Spread Voxel Pooling for Bird's-Eye-View Representation in Vision-based Roadside 3D Object Detection
Wenjie Wang, Yehao Lu, Guangcong Zheng, Shuigen Zhan, Xiaoqing Ye, Zichang Tan, Jingdong Wang, Gaoang Wang, Xi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14718-14727
Vision-based roadside 3D object detection has attracted rising attention in autonomous driving domain since it encompasses inherent advantages in reducing blind spots and expanding perception range. While previous work mainly focuses on accurately estimating depth or height for 2D-to-3D mapping ignoring the position approximation error in the voxel pooling process. Inspired by this insight we propose a novel voxel pooling strategy to reduce such error dubbed BEVSpread. Specifically instead of bringing the image features contained in a frustum point to a single BEV grid BEVSpread considers each frustum point as a source and spreads the image features to the surrounding BEV grids with adaptive weights. To achieve superior propagation performance a specific weight function is designed to dynamically control the decay speed of the weights according to distance and depth. Aided by customized CUDA parallel acceleration BEVSpread achieves comparable inference time as the original voxel pooling. Extensive experiments on two large-scale roadside benchmarks demonstrate that as a plug-in BEVSpread can significantly improve the performance of existing frustum-based BEV methods by a large margin of (1.12 5.26 3.01) AP in vehicle pedestrian and cyclist.
****
Real-IAD: A Real-World Multi-View Dataset for Benchmarking Versatile Industrial Anomaly Detection
Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, Lizhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22883-22892
Industrial anomaly detection (IAD) has garnered significant attention and experienced rapid development. However the recent development of IAD approach has enco

untered certain difficulties due to dataset limitations. On the one hand most of the state-of-the-art methods have achieved saturation (over 99% in AUROC) on mainstream datasets such as MVTec and the differences of methods cannot be well distinguished leading to a significant gap between public datasets and actual application scenarios. On the other hand the research on various new practical anomaly detection settings is limited by the scale of the dataset posing a risk of overfitting in evaluation results. Therefore we propose a large-scale Real-world and multi-view Industrial Anomaly Detection dataset named Real-IAD which contains 150K high-resolution images of 30 different objects an order of magnitude larger than existing datasets. It has a larger range of defect area and ratio proportions making it more challenging than previous datasets. To make the dataset closer to real application scenarios we adopted a multi-view shooting method and proposed sample-level evaluation metrics. In addition beyond the general unsupervised anomaly detection setting we propose a new setting for Fully Unsupervised Industrial Anomaly Detection (FUIAD) based on the observation that the yield rate in industrial production is usually greater than 60% which has more practical application value. Finally we report the results of popular IAD methods on the Real-IAD dataset providing a highly challenging benchmark to promote the development of the IAD field.
****
PAIR Diffusion: A Comprehensive Multimodal Object-Level Image Editor
Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, Humphrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8609-8618
Generative image editing has recently witnessed extremely fast-paced growth. Some works use high-level conditioning such as text while others use low-level conditioning. Nevertheless most of them lack fine-grained control over the properties of the different objects present in the image i.e. object-level image editing. In this work we tackle the task by perceiving the images as an amalgamation of various objects and aim to control the properties of each object in a fine-grained manner. Out of these properties we identify structure and appearance as the most intuitive to understand and useful for editing purposes. We propose PAIR Diffusion a generic framework that enables a diffusion model to control the structure and appearance properties of each object in the image. We show that having control over the properties of each object in an image leads to comprehensive editing capabilities. Our framework allows for various object-level editing operations on real images such as reference image-based appearance editing free-form shape editing adding objects and variations. Thanks to our design we do not require any inversion step. Additionally we propose multimodal classifier-free guidance which enables editing images using both reference images and text when using our approach with foundational diffusion models. We validate the above claims by extensively evaluating our framework on both unconditional and foundational diffusion models.
****
Boosting Adversarial Transferability by Block Shuffle and Rotation
Kunyu Wang, Xuanran He, Wenxuan Wang, Xiaosen Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24336-24346
Adversarial examples mislead deep neural networks with imperceptible perturbations and have brought significant threats to deep learning. An important aspect is their transferability which refers to their ability to deceive other models thus enabling attacks in the black-box setting. Though various methods have been proposed to boost transferability the performance still falls short compared with white-box attacks. In this work we observe that existing input transformation based attacks one of the mainstream transfer-based attacks result in different attention heatmaps on various models which might limit the transferability. We also find that breaking the intrinsic relation of the image can disrupt the attention heatmap of the original image. Based on this finding we propose a novel input transformation based attack called block shuffle and rotation (BSR). Specifically BSR splits the input image into several blocks then randomly shuffles and rota

tes these blocks to construct a set of new images for gradient calculation. Empi rical evaluations on the ImageNet dataset demonstrate that BSR could achieve sig nificantly better transferability than the existing input transformation based m ethods under single-model and ensemble-model settings. Combining BSR with the cu rrent input transformation method can further improve the transferability which significantly outperforms the state-of-the-art methods. Code is available at htt ps://github.com/Trustworthy-AI-Group/BSR.
****
DriveWorld: 4D Pre-trained Scene Understanding via World Models for Autonomous D riving
Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jians hu Li, Yulan Guo, Junliang Xing, Liping Jing, Yiming Nie, Bin Dai; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 15522-15533
Vision-centric autonomous driving has recently raised wide attention due to its lower cost. Pre-training is essential for extracting a universal representation. However current vision-centric pre-training typically relies on either 2D or 3D pre-text tasks overlooking the temporal characteristics of autonomous driving a s a 4D scene understanding task. In this paper we address this challenge by intr oducing a world model-based autonomous driving 4D representation learning framew ork dubbed DriveWorld which is capable of pre-training from multi-camera driving videos in a spatio-temporal fashion. Specifically we propose a Memory State-Spa ce Model for spatio-temporal modelling which consists of a Dynamic Memory Bank m odule for learning temporal-aware latent dynamics to predict future changes and a Static Scene Propagation module for learning spatial-aware latent statics to o ffer comprehensive scene contexts. We additionally introduce a Task Prompt to de couple task-aware features for various downstream tasks. The experiments demonst rate that DriveWorld delivers promising results on various autonomous driving ta sks. When pre-trained with the OpenScene dataset DriveWorld achieves a 7.5% incr ease in mAP for 3D object detection a 3.0% increase in IoU for online mapping a 5.0% increase in AMOTA for multi-object tracking a 0.1m decrease in minADE for m otion forecasting a 3.0% increase in IoU for occupancy prediction and a 0.34m re duction in average L2 error for planning.
****
Bridging the Gap Between End-to-End and Two-Step Text Spotting
Mingxin Huang, Hongliang Li, Yuliang Liu, Xiang Bai, Lianwen Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024 , pp. 15608-15618
Modularity plays a crucial role in the development and maintenance of complex sy stems. While end-to-end text spotting efficiently mitigates the issues of error accumulation and sub-optimal performance seen in traditional two-step methodolog ies the two-step methods continue to be favored in many competitions and practic al settings due to their superior modularity. In this paper we introduce Bridgin g Text Spotting a novel approach that resolves the error accumulation and subopt imal performance issues in two-step methods while retaining modularity. To achie ve this we adopt a well-trained detector and recognizer that are developed and t rained independently and then lock their parameters to preserve their already ac quired capabilities. Subsequently we introduce a Bridge that connects the locked detector and recognizer through a zero-initialized neural network. This zero-in itialized neural network initialized with weights set to zeros ensures seamless integration of the large receptive field features in detection into the locked r ecognizer. Furthermore since the fixed detector and recognizer cannot naturally acquire end-to-end optimization features we adopt the Adapter to facilitate thei r efficient learning of these features. We demonstrate the effectiveness of the proposed method through extensive experiments: Connecting the latest detector an d recognizer through Bridging Text Spotting we achieved an accuracy of 83.3% on Total-Text 69.8% on CTW1500 and 89.5% on ICDAR 2015. The code is available at ht tps://github.com/mxin262/Bridging-Text-Spotting.
****
TokenCompose: Text-to-Image Diffusion with Token-level Supervision

Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, Zhuowen Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8553-8564
We present TokenCompose a Latent Diffusion Model for text-to-image generation that achieves enhanced consistency between user-specified text prompts and model-generated images. Despite its tremendous success the standard denoising process in the Latent Diffusion Model takes text prompts as conditions only absent explicit constraint for the consistency between the text prompts and the image contents leading to unsatisfactory results for composing multiple object categories. Our proposed TokenCompose aims to improve multi-category instance composition by introducing the token-wise consistency terms between the image content and object segmentation maps in the finetuning stage. TokenCompose can be applied directly to the existing training pipeline of text-conditioned diffusion models without extra human labeling information. By finetuning Stable Diffusion with our approach the model exhibits significant improvements in multi-category instance composition and enhanced photorealism for its generated images.
****
SUGAR: Pre-training 3D Visual Representations for Robotics
Shizhe Chen, Ricardo Garcia, Ivan Laptev, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18049-18060
Learning generalizable visual representations from Internet data has yielded promising results for robotics. Yet prevailing approaches focus on pre-training 2D representations being sub-optimal to deal with occlusions and accurately localize objects in complex 3D scenes. Meanwhile 3D representation learning has been limited to single-object understanding. To address these limitations we introduce a novel 3D pre-training framework for robotics named SUGAR that captures semantic geometric and affordance properties of objects through 3D point clouds. We underscore the importance of cluttered scenes in 3D representation learning and automatically construct a multi-object dataset benefiting from cost-free supervision in simulation. SUGAR employs a versatile transformer-based model to jointly address five pre-training tasks namely cross-modal knowledge distillation for semantic learning masked point modeling to understand geometry structures grasping pose synthesis for object affordance 3D instance segmentation and referring expression grounding to analyze cluttered scenes. We evaluate our learned representation on three robotic-related tasks namely zero-shot 3D object recognition referring expression grounding and language-driven robotic manipulation. Experimental results show that SUGAR's 3D representation outperforms state-of-the-art 2D and 3D representations.
****
LidaRF: Delving into Lidar for Neural Radiance Field on Street Scenes
Shanlin Sun, Bingbing Zhuang, Ziyu Jiang, Buyu Liu, Xiaohui Xie, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19563-19572
Photorealistic simulation plays a crucial role in applications such as autonomous driving where advances in neural radiance fields (NeRFs) may allow better scalability through the automatic creation of digital 3D assets. However reconstruction quality suffers on street scenes due to largely collinear camera motions and sparser samplings at higher speeds. On the other hand the application often demands rendering from camera views that deviate from the inputs to accurately simulate behaviors like lane changes. In this paper we propose several insights that allow a better utilization of Lidar data to improve NeRF quality on street scenes. First our framework learns a geometric scene representation from Lidar which are fused with the implicit grid-based representation for radiance decoding thereby supplying stronger geometric information offered by explicit point cloud. Second we put forth a robust occlusion-aware depth supervision scheme which allows utilizing densified Lidar points by accumulation. Third we generate augmented training views from Lidar points for further improvement. Our insights translate to largely improved novel view synthesis under real driving scenes.
****

PairAug: What Can Augmented Image-Text Pairs Do for Radiology?
Yutong Xie, Qi Chen, Sinuo Wang, Minh-Son To, Iris Lee, Ee Win Khoo, Kerolos Hendy, Daniel Koh, Yong Xia, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11652-11661
Current vision-language pre-training (VLP) methodologies predominantly depend on paired image-text datasets a resource that is challenging to acquire in radiology due to privacy considerations and labelling complexities. Data augmentation provides a practical solution to overcome the issue of data scarcity however most augmentation methods exhibit a limited focus prioritising either image or text augmentation exclusively. Acknowledging this limitation our objective is to devise a framework capable of concurrently augmenting medical image and text data. We design a Pairwise Augmentation (PairAug) approach that contains an Inter-patient Augmentation (InterAug) branch and an Intra-patient Augmentation (IntraAug) branch. Specifically the InterAug branch of our approach generates radiology images using synthesised yet plausible reports derived from a Large Language Model (LLM). The generated pairs can be considered a collection of new patient cases since they are artificially created and may not exist in the original dataset. In contrast the IntraAug branch uses newly generated reports to manipulate images. This process allows us to create new paired data for each individual with diverse medical conditions. Our extensive experiments on various downstream tasks covering medical image classification zero-shot and fine-tuning analysis demonstrate that our PairAug concurrently expanding both image and text data substantially outperforms image-/text-only expansion baselines and advanced medical VLP baselines. Our code is released at https://github.com/YtongXie/PairAug.
****
FINER: Flexible Spectral-bias Tuning in Implicit NEural Representation by Variable-periodic Activation Functions
Zhen Liu, Hao Zhu, Qi Zhang, Jingde Fu, Weibing Deng, Zhan Ma, Yanwen Guo, Xun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2713-2722
Implicit Neural Representation (INR) which utilizes a neural network to map coordinate inputs to corresponding attributes is causing a revolution in the field of signal processing. However current INR techniques suffer from a restricted capability to tune their supported frequency set resulting in imperfect performance when representing complex signals with multiple frequencies. We have identified that this frequency-related problem can be greatly alleviated by introducing variable-periodic activation functions for which we propose FINER. By initializing the bias of the neural network within different ranges sub-functions with various frequencies in the variable-periodic function are selected for activation. Consequently the supported frequency set of FINER can be flexibly tuned leading to improved performance in signal representation. We demonstrate the capabilities of FINER in the contexts of 2D image fitting 3D signed distance field representation and 5D neural radiance fields optimization and we show that it outperforms existing INRs.
****
Harnessing Large Language Models for Training-free Video Anomaly Detection
Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18527-18536
Video anomaly detection (VAD) aims to temporally locate abnormal events in a video. Existing works mostly rely on training deep models to learn the distribution of normality with either video-level supervision one-class supervision or in an unsupervised setting. Training-based methods are prone to be domain-specific thus being costly for practical deployment as any domain change will involve data collection and model training. In this paper we radically depart from previous efforts and propose LAnguage-based VAD (LAVAD) a method tackling VAD in a novel training-free paradigm exploiting the capabilities of pre-trained large language models (LLMs) and existing vision-language models (VLMs). We leverage VLM-based captioning models to generate textual descriptions for each frame of any test video. With the textual scene description we then devise a prompting mechanism to

unlock the capability of LLMs in terms of temporal aggregation and anomaly score
 estimation turning LLMs into an effective video anomaly detector. We further le
verage modality-aligned VLMs and propose effective techniques based on cross-mod
al similarity for cleaning noisy captions and refining the LLM-based anomaly sco
res. We evaluate LAVAD on two large datasets featuring real-world surveillance s
cenarios (UCF-Crime and XD-Violence) showing that it outperforms both unsupervis
ed and one-class methods without requiring any training or data collection.
****
TextCraftor: Your Text Encoder Can be Image Quality Controller
Yanyu Li, Xian Liu, Anil Kag, Ju Hu, Yerlan Idelbayev, Dhritiman Sagar, Yanzhi W
ang, Sergey Tulyakov, Jian Ren; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2024, pp. 7985-7995
Diffusion-based text-to-image generative models e.g. Stable Diffusion have revol
utionized the field of content generation enabling significant advancements in a
reas like image editing and video synthesis. Despite their formidable capabiliti
es these models are not without their limitations. It is still challenging to sy
nthesize an image that aligns well with the input text and multiple runs with ca
refully crafted prompts are required to achieve satisfactory results. To mitigat
e these limitations numerous studies have endeavored to fine-tune the pre-traine
d diffusion models i.e.. UNet utilizing various technologies. Yet amidst these e
fforts a pivotal question of text-to-image diffusion model training has remained
 largely unexplored: Is it possible and feasible to fine-tune the text encoder t
o improve the performance of text-to-image diffusion models? Our findings reveal
 that instead of replacing the CLIP text encoder used in Stable Diffusion with o
ther large language models we can enhance it through our proposed fine-tuning ap
proach TextCraftor leading to substantial improvements in quantitative benchmark
s and human assessments. Interestingly our technique also empowers controllable
image generation through the interpolation of different text encoders fine-tuned
 with various rewards. We also demonstrate that TextCraftor is orthogonal to UNe
t finetuning and can be combined to further improve generative quality.
****
FineParser: A Fine-grained Spatio-temporal Action Parser for Human-centric Actio
n Quality Assessment
Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, Yuxin Peng; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
14628-14637
Existing action quality assessment (AQA) methods mainly learn deep representatio
ns at the video level for scoring diverse actions. Due to the lack of a fine-gra
ined understanding of actions in videos they harshly suffer from low credibility
 and interpretability thus insufficient for stringent applications such as Olymp
ic diving events. We argue that a fine-grained understanding of actions requires
 the model to perceive and parse actions in both time and space which is also th
e key to the credibility and interpretability of the AQA technique. Based on thi
s insight we propose a new fine-grained spatial-temporal action parser named Fin
eParser. It learns human-centric foreground action representations by focusing o
n target action regions within each frame and exploiting their fine-grained alig
nments in time and space to minimize the impact of invalid backgrounds during th
e assessment. In addition we construct fine-grained annotations of human-centric
 foreground action masks for the FineDiving dataset called FineDiving-HM. With r
efined annotations on diverse target action procedures FineDiving-HM can promote
 the development of real-world AQA systems. Through extensive experiments we dem
onstrate the effectiveness of FineParser which outperforms state-of-the-art meth
ods while supporting more tasks of fine-grained action understanding. Data and c
ode are available at https://github.com/PKU-ICST-MIPL/FineParser_CVPR2024.
****
Video Recognition in Portrait Mode
Mingfei Han, Linjie Yang, Xiaojie Jin, Jiashi Feng, Xiaojun Chang, Heng Wang; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 21831-21841
The creation of new datasets often presents new challenges for video recognition

 and can inspire novel ideas while addressing these challenges. While existing d
atasets mainly comprise landscape mode videos our paper seeks to introduce portr
ait mode videos to the research community and highlight the unique challenges as
sociated with this video format. With the growing popularity of smartphones and
social media applications recognizing portrait mode videos is becoming increasin
gly important. To this end we have developed the first dataset dedicated to port
rait mode video recognition namely PortraitMode-400. The taxonomy of PortraitMod
e-400 was constructed in a data-driven manner comprising 400 fine-grained catego
ries and rigorous quality assurance was implemented to ensure the accuracy of hu
man annotations. In addition to the new dataset we conducted a comprehensive ana
lysis of the impact of video format (portrait mode versus landscape mode) on rec
ognition accuracy and spatial bias due to the different formats. Furthermore we
designed extensive experiments to explore key aspects of portrait mode video rec
ognition including the choice of data augmentation evaluation procedure the impo
rtance of temporal information and the role of audio modality. Building on the i
nsights from our experimental results and the introduction of PortraitMode-400 o
ur paper aims to inspire further research efforts in this emerging research area
.
****
Selective Hourglass Mapping for Universal Image Restoration Based on Diffusion M
odel
Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, Wei-Shi Zheng;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2024, pp. 25445-25455
Universal image restoration is a practical and potential computer vision task fo
r real-world applications. The main challenge of this task is handling the diffe
rent degradation distributions at once. Existing methods mainly utilize task-spe
cific conditions (e.g. prompt) to guide the model to learn different distributio
ns separately named multi-partite mapping. However it is not suitable for univer
sal model learning as it ignores the shared information between different tasks.
 In this work we propose an advanced selective hourglass mapping strategy based
on diffusion model termed DiffUIR. Two novel considerations make our DiffUIR non
-trivial. Firstly we equip the model with strong condition guidance to obtain ac
curate generation direction of diffusion model (selective). More importantly Dif
fUIR integrates a flexible shared distribution term (SDT) into the diffusion alg
orithm elegantly and naturally which gradually maps different distributions into
 a shared one. In the reverse process combined with SDT and strong condition gui
dance DiffUIR iteratively guides the shared distribution to the task-specific di
stribution with high image quality (hourglass). Without bells and whistles by on
ly modifying the mapping strategy we achieve state-of-the-art performance on fiv
e image restoration tasks 22 benchmarks in the universal setting and zero-shot g
eneralization setting. Surprisingly by only using a lightweight model (only 0.89
M) we could achieve outstanding performance. The source code and pre-trained mod
els are available at https://github.com/iSEE-Laboratory/DiffUIR
****
Language Models as Black-Box Optimizers for Vision-Language Models
Shihong Liu, Samuel Yu, Zhiqiu Lin, Deepak Pathak, Deva Ramanan; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 12687-12697
Vision-language models (VLMs) pre-trained on web-scale datasets have demonstrate
d remarkable capabilities on downstream tasks when fine-tuned with minimal data.
 However many VLMs rely on proprietary data and are not open-source which restri
cts the use of white-box approaches for fine-tuning. As such we aim to develop a
 black-box approach to optimize VLMs through natural language prompts thereby
avoiding the need to access model parameters feature embeddings or even output l
ogits. We propose employing chat-based LLMs to search for the best text prompt f
or VLMs. Specifically we adopt an automatic "hill-climbing" procedure that conve
rges to an effective prompt by evaluating the performance of current prompts and
 asking LLMs to refine them based on textual feedback all within a conversationa
l process without human-in-the-loop. In a challenging 1-shot image classificatio

n setup our simple approach surpasses the white-box continuous prompting method (CoOp) by an average of 1.5% across 11 datasets including ImageNet. Our approach also outperforms both human-engineered and LLM-generated prompts. We highlight the advantage of conversational feedback that incorporates both positive and negative prompts suggesting that LLMs can utilize the implicit "gradient" direction in textual feedback for a more efficient search. In addition we find that the text prompts generated through our strategy are not only more interpretable but also transfer well across different VLM architectures in a black-box manner. Lastly we demonstrate our framework on a state-of-the-art black-box VLM (DALL-E 3) for text-to-image optimization.
****
Exploring Orthogonality in Open World Object Detection
Zhicheng Sun, Jinghan Li, Yadong Mu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17302-17312
Open world object detection aims to identify objects of unseen categories and incrementally recognize them once their annotations are provided. In distinction to the traditional paradigm that is limited to predefined categories this setting promises a continual and generalizable way of estimating objectness using class-agnostic information. However achieving such decorrelation between objectness and class information proves challenging. Without explicit consideration existing methods usually exhibit low recall on unknown objects and can misclassify them into known classes. To address this problem we exploit three levels of orthogonality in the detection process: First the objectness and classification heads are disentangled by operating on separate sets of features that are orthogonal to each other in a devised polar coordinate system. Secondly a prediction decorrelation loss is introduced to guide the detector towards more general and class-independent prediction. Furthermore we propose a calibration scheme that helps maintain orthogonality throughout the training process to mitigate catastrophic interference and facilitate incremental learning of previously unseen objects. Our method is comprehensively evaluated on open world and incremental object detection benchmarks demonstrating its effectiveness in detecting both known and unknown objects. Code and models are available at https://github.com/feifeiobama/OrthogonalDet.
****
Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding
Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, Lidong Bing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13872-13882
Large Vision-Language Models (LVLMs) have advanced considerably intertwining visual recognition and language understanding to generate content that is not only coherent but also contextually attuned. Despite their success LVLMs still suffer from the issue of object hallucinations where models generate plausible yet incorrect outputs that include objects that do not exist in the images. To mitigate this issue we introduce Visual Contrastive Decoding (VCD) a simple and training-free method that contrasts output distributions derived from original and distorted visual inputs. The proposed VCD effectively reduces the over-reliance on statistical bias and unimodal priors two essential causes of object hallucinations. This adjustment ensures the generated content is closely grounded to visual inputs resulting in contextually accurate outputs. Our experiments show that VCD without either additional training or the usage of external tools significantly mitigates the object hallucination issue across different LVLM families. Beyond mitigating object hallucinations VCD also excels in general LVLM benchmarks highlighting its wide-ranging applicability.
****
IMPRINT: Generative Object Compositing by Learning Identity-Preserving Representation
Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, Daniel Aliaga; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8048-8058

Generative object compositing emerges as a promising new avenue for compositiona
l image editing. However the requirement of object identity preservation poses a
 significant challenge limiting practical usage of most existing methods. In res
ponse this paper introduces IMPRINT a novel diffusion-based generative model tra
ined with a two-stage learning framework that decouples learning of identity pre
servation from that of compositing. The first stage is targeted for context-agno
stic identity-preserving pretraining of the object encoder enabling the encoder
to learn an embedding that is both view-invariant and conducive to enhanced deta
il preservation. The subsequent stage leverages this representation to learn sea
mless harmonization of the object composited to the background. In addition IMPR
INT incorporates a shape-guidance mechanism offering user-directed control over
the compositing process. Extensive experiments demonstrate that IMPRINT signific
antly outperforms existing methods and various baselines on identity preservatio
n and composition quality.
****
Audio-Visual Segmentation via Unlabeled Frame Exploitation
Jinxiang Liu, Yikun Liu, Fei Zhang, Chen Ju, Ya Zhang, Yanfeng Wang; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 26328-26339
Audio-visual segmentation (AVS) aims to segment the sounding objects in video fr
ames. Although great progress has been witnessed we experimentally reveal that c
urrent methods reach marginal performance gain within the use of the unlabeled f
rames leading to the underutilization issue. To fully explore the potential of t
he unlabeled frames for AVS we explicitly divide them into two categories based
on their temporal characteristics i.e. neighboring frame (NF) and distant frame
(DF). NFs temporally adjacent to the labeled frame often contain rich motion inf
ormation that assists in the accurate localization of sounding objects. Contrary
 to NFs DFs have long temporal distances from the labeled frame which share sema
ntic-similar objects with appearance variations. Considering their unique charac
teristics we propose a versatile framework that effectively leverages them to ta
ckle AVS. Specifically for NFs we exploit the motion cues as the dynamic guidanc
e to improve the objectness localization. Besides we exploit the semantic cues i
n DFs by treating them as valid augmentations to the labeled frames which are th
en used to enrich data diversity in a self-training manner. Extensive experiment
al results demonstrate the versatility and superiority of our method unleashing
the power of the abundant unlabeled frames.
****
DriveTrack: A Benchmark for Long-Range Point Tracking in Real-World Videos
Arjun Balasingam, Joseph Chandler, Chenning Li, Zhoutong Zhang, Hari Balakrishna
n; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2024, pp. 22488-22497
This paper presents DriveTrack a new benchmark and data generation framework for
 long-range keypoint tracking in real-world videos. DriveTrack is motivated by t
he observation that the accuracy of state-of-the-art trackers depends strongly o
n visual attributes around the selected keypoints such as texture and lighting.
The problem is that these artifacts are especially pronounced in real-world vide
os but these trackers are unable to train on such scenes due to a dearth of anno
tations. DriveTrack bridges this gap by building a framework to automatically an
notate point tracks on autonomous driving datasets. We release a dataset consist
ing of 1 billion point tracks across 24 hours of video which is seven orders of
magnitude greater than prior real-world benchmarks and on par with the scale of
synthetic benchmarks. DriveTrack unlocks new use cases for point tracking in rea
l-world videos. First we show that fine-tuning keypoint trackers on DriveTrack i
mproves accuracy on real-world scenes by up to 7%. Second we analyze the sensiti
vity of trackers to visual artifacts in real scenes and motivate the idea of run
ning assistive keypoint selectors alongside trackers.
****
Infrared Adversarial Car Stickers
Xiaopei Zhu, Yuqiu Liu, Zhanhao Hu, Jianmin Li, Xiaolin Hu; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.

24284-24293
Infrared physical adversarial examples are of great significance for studying th
e security of infrared AI systems that are widely used in our lives such as auto
nomous driving. Previous infrared physical attacks mainly focused on 2D infrared
 pedestrian detection which may not fully manifest its destructiveness to AI sys
tems. In this work we propose a physical attack method against infrared detector
s based on 3D modeling which is applied to a real car. The goal is to design a s
et of infrared adversarial stickers to make cars invisible to infrared detectors
 at various viewing angles distances and scenes. We build a 3D infrared car mode
l with real infrared characteristics and propose an infrared adversarial pattern
 generation method based on 3D mesh shadow. We propose a 3D control points-based
 mesh smoothing algorithm and use a set of smoothness loss functions to enhance
the smoothness of adversarial meshes and facilitate the sticker implementation.
Besides We designed the aluminum stickers and conducted physical experiments on
two real Mercedes-Benz A200L cars. Our adversarial stickers hid the cars from Fa
ster RCNN an object detector at various viewing angles distances and scenes. The
 attack success rate (ASR) was 91.49% for real cars. In comparison the ASRs of r
andom stickers and no sticker were only 6.21% and 0.66% respectively. In additio
n the ASRs of the designed stickers against six unseen object detectors such as
YOLOv3 and Deformable DETR were between 73.35%-95.80% showing good transferabili
ty of the attack performance across detectors.
****
Sculpt3D: Multi-View Consistent Text-to-3D Generation with Sparse 3D Prior
Cheng Chen, Xiaofeng Yang, Fan Yang, Chengzeng Feng, Zhoujie Fu, Chuan-Sheng Foo
, Guosheng Lin, Fayao Liu; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2024, pp. 10228-10237
Recent works on text-to-3d generation show that using only 2D diffusion supervis
ion for 3D generation tends to produce results with inconsistent appearances (e.
g. faces on the back view) and inaccurate shapes (e.g. animals with extra legs).
 Existing methods mainly address this issue by retraining diffusion models with
images rendered from 3D data to ensure multi-view consistency while struggling t
o balance 2D generation quality with 3D consistency. In this paper we present a
new framework Sculpt3D that equips the current pipeline with explicit injection
of 3D priors from retrieved reference objects without re-training the 2D diffusi
on model. Specifically we demonstrate that high-quality and diverse 3D geometry
can be guaranteed by keypoints supervision through a sparse ray sampling approac
h. Moreover to ensure accurate appearances of different views we further modulat
e the output of the 2D diffusion model to the correct patterns of the template v
iews without altering the generated object's style. These two decoupled designs
effectively harness 3D information from reference objects to generate 3D objects
 while preserving the generation quality of the 2D diffusion model. Extensive ex
periments show our method can largely improve the multi-view consistency while r
etaining fidelity and diversity.
****
FreeMan: Towards Benchmarking 3D Human Pose Estimation under Real-World Conditio
ns
Jiong Wang, Fengyu Yang, Bingliang Li, Wenbo Gou, Danqi Yan, Ailing Zeng, Yijun
Gao, Junle Wang, Yanqing Jing, Ruimao Zhang; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21978-21988
Estimating the 3D structure of the human body from nat- ural scenes is a fundame
ntal aspect of visual perception. 3D human pose estimation is a vital step in ad
vancing fields like AIGC and human-robot interaction serving as a crucial tech-
nique for understanding and interacting with human actions in real-world setting
s. However the current datasets often collected under single laboratory conditio
ns using complex motion capture equipment and unvarying backgrounds are insuffic
ient. The absence of datasets on variable conditions is stalling the progress of
 this crucial task. To facilitate the development of 3D pose estimation we prese
nt FreeMan the first large-scale multi-view dataset collected under the real- wo
rld conditions. FreeMan was captured by synchronizing 8 smartphones across diver
se scenarios. It comprises 11M frames from 8000 sequences viewed from different

perspec- tives. These sequences cover 40 subjects across 10 different scenarios each with varying lighting conditions. We have also established an semi-automated pipeline containing er- ror detection to reduce the workload of manual check and ensure precise annotation. We provide comprehensive eval- uation baselines for a range of tasks underlining the sig- nificant challenges posed by FreeMan. Further evaluations of standard indoor/outdoor human sensing datasets reveal that FreeMan offers robust representation transferability in real and complex scenes. FreeMan is publicly available at https://wangjiongw.github.io/freeman.
****
ScanFormer: Referring Expression Comprehension by Iteratively Scanning
Wei Su, Peihan Miao, Huanzhang Dou, Xi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13449-13458
Referring Expression Comprehension (REC) aims to localize the target objects specified by free-form natural language descriptions in images. While state-of-the-art methods achieve impressive performance they perform a dense perception of images which incorporates redundant visual regions unrelated to linguistic queries leading to additional computational overhead. This inspires us to explore a question: can we eliminate linguistic-irrelevant redundant visual regions to improve the efficiency of the model? Existing relevant methods primarily focus on fundamental visual tasks with limited exploration in vision-language fields. To address this we propose a coarse-to-fine iterative perception framework called ScanFormer. It can iteratively exploit the image scale pyramid to extract linguistic-relevant visual patches from top to bottom. In each iteration irrelevant patches are discarded by our designed informativeness prediction. Furthermore we propose a patch selection strategy for discarded patches to accelerate inference. Experiments on widely used datasets namely RefCOCO RefCOCO+ RefCOCOg and ReferItGame verify the effectiveness of our method which can strike a balance between accuracy and efficiency.
****
Model Inversion Robustness: Can Transfer Learning Help?
Sy-Tuyen Ho, Koh Jun Hao, Keshigeyan Chandrasegaran, Ngoc-Bao Nguyen, Ngai-Man Cheung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12183-12193
Model Inversion (MI) attacks aim to reconstruct private training data by abusing access to machine learning models. Contemporary MI attacks have achieved impressive attack performance posing serious threats to privacy. Meanwhile all existing MI defense methods rely on regularization that is in direct conflict with the training objective resulting in noticeable degradation in model utility. In this work we take a different perspective and propose a novel and simple Transfer Learning-based Defense against Model Inversion (TL-DMI) to render MI-robust models. Particularly by leveraging TL we limit the number of layers encoding sensitive information from private training dataset thereby degrading the performance of MI attack. We conduct an analysis using Fisher Information to justify our method. Our defense is remarkably simple to implement. Without bells and whistles we show in extensive experiments that TL-DMI achieves state-of-the-art (SOTA) MI robustness. Our code pre-trained models demo and inverted data are available at: https://hosytuyen.github.io/projects/TL-DMI
****
Portrait4D: Learning One-Shot 4D Head Avatar Synthesis using Synthetic Data
Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, Baoyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7119-7130
Existing one-shot 4D head synthesis methods usually learn from monocular videos with the aid of 3DMM reconstruction yet the latter is evenly challenging which restricts them from reasonable 4D head synthesis. We present a method to learn one-shot 4D head synthesis via large-scale synthetic data. The key is to first learn a part-wise 4D generative model from monocular images via adversarial learning to synthesize multi-view images of diverse identities and full motions as training data; then leverage a transformer-based animatable triplane reconstructor to learn 4D head reconstruction using the synthetic data. A novel learning strate

gy is enforced to enhance the generalizability to real images by disentangling the learning process of 3D reconstruction and reenactment. Experiments demonstrate our superiority over the prior art.
****
GP-NeRF: Generalized Perception NeRF for Context-Aware 3D Scene Understanding
Hao Li, Dingwen Zhang, Yalun Dai, Nian Liu, Lechao Cheng, Jingfeng Li, Jingdong Wang, Junwei Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21708-21718
Applying Neural Radiance Fields (NeRF) to downstream perception tasks for scene understanding and representation is becoming increasingly popular. Most existing methods treat semantic prediction as an additional rendering task i.e. the "label rendering" task to build semantic NeRFs. However by rendering semantic/instance labels per pixel without considering the contextual information of the rendered image these methods usually suffer from unclear boundary segmentation and abnormal segmentation of pixels within an object. To solve this problem we propose Generalized Perception NeRF (GP-NeRF) a novel pipeline that makes the widely used segmentation model and NeRF work compatibly under a unified framework for facilitating context-aware 3D scene perception. To accomplish this goal we introduce transformers to aggregate radiance as well as semantic embedding fields jointly for novel views and facilitate the joint volumetric rendering of both fields. In addition we propose two self-distillation mechanisms i.e. the Semantic Distill Loss and the Depth-Guided Semantic Distill Loss to enhance the discrimination and quality of the semantic field and the maintenance of geometric consistency. In evaluation as shown in Fig. 1 we conduct experimental comparisons under two perception tasks (i.e. semantic and instance segmentation) using both synthetic and real-world datasets. Notably our method outperforms SOTA approaches by 6.94% 11.76% and 8.47% on generalized semantic segmentation finetuning semantic segmentation and instance segmentation respectively
****
Polarization Wavefront Lidar: Learning Large Scene Reconstruction from Polarized Wavefronts
Dominik Scheuble, Chenyang Lei, Seung-Hwan Baek, Mario Bijelic, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21241-21250
Lidar has become a cornerstone sensing modality for 3D vision especially for large outdoor scenarios and autonomous driving. Conventional lidar sensors are capable of providing centimeter-accurate distance information by emitting laser pulses into a scene and measuring the time-of-flight (ToF) of the reflection. However the polarization of the received light that depends on the surface orientation and material properties is usually not considered. As such the polarization modality has the potential to improve scene reconstruction beyond distance measurements. In this work we introduce a novel long-range polarization wavefront lidar sensor (PolLidar) that modulates the polarization of the emitted and received light. Departing from conventional lidar sensors PolLidar allows access to the raw time-resolved polarimetric wavefronts. We leverage polarimetric wavefronts to estimate normals distance and material properties in outdoor scenarios with a novel learned reconstruction method. To train and evaluate the method we introduce a simulated and real-world long-range dataset with paired raw lidar data ground truth distance and normal maps. We find that the proposed method improves normal and distance reconstruction by 53% mean angular error and 41% mean absolute error compared to existing shape-from-polarization (SfP) and ToF methods. Code and data are open-sourced here.
****
GDA: Generalized Diffusion for Robust Test-time Adaptation
Yun-Yun Tsai, Fu-Chen Chen, Albert Y. C. Chen, Junfeng Yang, Che-Chun Su, Min Sun, Cheng-Hao Kuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23242-23251
Machine learning models face generalization challenges when exposed to out-of-distribution (OOD) samples with unforeseen distribution shifts. Recent research reveals that for vision tasks test-time adaptation employing diffusion models can

achieve state-of-the-art accuracy improvements on OOD samples by generating doma in-aligned samples without altering the model's weights. Unfortunately those stu dies have primarily focused on pixel-level corruptions thereby lacking the gener alization to adapt to a broader range of OOD types. We introduce Generalized Dif fusion Adaptation (GDA) a novel diffusion-based test-time adaptation method robu st against diverse OOD types. Specifically GDA iteratively guides the diffusion by applying a marginal entropy loss derived from the model in conjunction with s tyle and content preservation losses during the reverse sampling process. In oth er words GDA considers the model's output behavior and the samples' semantic inf ormation as a whole reducing ambiguity in downstream tasks. based adaptation. Ev aluation across various model architectures and OOD benchmarks indicates that GD A consistently surpasses previous diffusion-based adaptation methods. Notably it achieves the highest classification accuracy improvements ranging from 4.4% to 5.02% on ImageNet-C and 2.5% to 7.4% on Rendition Sketch and Stylized benchmarks . This performance highlights GDA's generalization to a broader range of OOD ben chmarks.
****

ConvoFusion: Multi-Modal Conversational Diffusion for Co-Speech Gesture Synthesi s

Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc H abermann, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1388-1398

Gestures play a key role in human communication. Recent methods for co-speech ge sture generation while managing to generate beat-aligned motions struggle genera ting gestures that are semantically aligned with the utterance. Compared to beat gestures that align naturally to the audio signal semantically coherent gesture s require modeling the complex interactions between the language and human motio n and can be controlled by focusing on certain words. Therefore we present Convo Fusion a diffusion-based approach for multi-modal gesture synthesis which can no t only generate gestures based on multi-modal speech inputs but can also facilit ate controllability in gesture synthesis. Our method proposes two guidance objec tives that allow the users to modulate the impact of different conditioning moda lities (e.g. audio vs text) as well as to choose certain words to be emphasized during gesturing. Our method is versatile in that it can be trained either for g enerating monologue gestures or even the conversational gestures. To further adv ance the research on multi-party interactive gestures the DnD Group Gesture data set is released which contains 6 hours of gesture data showing 5 people interact ing with one another. We compare our method with several recent works and demons trate effectiveness of our method on a variety of tasks. We urge the reader to w atch our supplementary video at https://vcai.mpi-inf.mpg.de/projects/ConvoFusion /.
****

RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Corre ctional Human Feedback

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zh iyuan Liu, Hai-Tao Zheng, Maosong Sun, Tat-Seng Chua; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13807- 13816

Multimodal Large Language Models (MLLMs) have recently demonstrated impressive c apabilities in multimodal understanding reasoning and interaction. However exist ing MLLMs prevalently suffer from serious hallucination problems generating text that is not factually grounded in associated images. The problem makes existing MLLMs untrustworthy and thus impractical in real-world (especially high-stakes) applications. To address the challenge we present RLHF-V which enhances MLLM tr ustworthiness via behavior alignment from fine-grained correctional human feedba ck. Specifically RLHF-V collects human preference in the form of segment-level c orrections on hallucinations and performs dense direct preference optimization o ver the human feedback. Comprehensive experiments on five benchmarks in both aut omatic and human evaluation show that RLHF-V can enable substantially more trust worthy MLLM behaviors with promising data and computation efficiency. Remarkably

using 1.4k annotated data samples RLHF-V significantly reduces the hallucination rate of the base MLLM by 34.8% outperforming the concurrent LLaVA-RLHF trained on 10k annotated data. The final model achieves state-of-the-art performance in trustworthiness among open-source MLLMs and shows better robustness than GPT-4V in preventing hallucinations aroused from over-generalization. All the data code and model weights will be released to facilitate future research.
****
ZeroShape: Regression-based Zero-shot Shape Reconstruction
Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, James M. Rehg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10061-10071
We study the problem of single-image zero-shot 3D shape reconstruction. Recent works learn zero-shot shape reconstruction through generative modeling of 3D assets but these models are computationally expensive at train and inference time. In contrast the traditional approach to this problem is regression-based where deterministic models are trained to directly regress the object shape. Such regression methods possess much higher computational efficiency than generative methods. This raises a natural question: is generative modeling necessary for high performance or conversely are regression-based approaches still competitive? To answer this we design a strong regression-based model called ZeroShape based on the converging findings in this field and a novel insight. We also curate a large real-world evaluation benchmark with objects from three different real-world 3D datasets. This evaluation benchmark is more diverse and an order of magnitude larger than what prior works use to quantitatively evaluate their models aiming at reducing the evaluation variance in our field. We show that ZeroShape not only achieves superior performance over state-of-the-art methods but also demonstrates significantly higher computational and data efficiency.
****
Continual-MAE: Adaptive Distribution Masked Autoencoders for Continual Test-Time Adaptation
Jiaming Liu, Ran Xu, Senqiao Yang, Renrui Zhang, Qizhe Zhang, Zehui Chen, Yandong Guo, Shanghang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28653-28663
Continual Test-Time Adaptation (CTTA) is proposed to migrate a source pre-trained model to continually changing target distributions addressing real-world dynamism. Existing CTTA methods mainly rely on entropy minimization or teacher-student pseudo-labeling schemes for knowledge extraction in unlabeled target domains. However dynamic data distributions cause miscalibrated predictions and noisy pseudo-labels in existing self-supervised learning methods hindering the effective mitigation of error accumulation and catastrophic forgetting problems during the continual adaptation process. To tackle these issues we propose a continual self-supervised method Adaptive Distribution Masked Autoencoders (ADMA) which enhances the extraction of target domain knowledge while mitigating the accumulation of distribution shifts. Specifically we propose a Distribution-aware Masking (DaM) mechanism to adaptively sample masked positions followed by establishing consistency constraints between the masked target samples and the original target samples. Additionally for masked tokens we utilize an efficient decoder to reconstruct a hand-crafted feature descriptor (e.g. Histograms of Oriented Gradients) leveraging its invariant properties to boost task-relevant representations. Through conducting extensive experiments on four widely recognized benchmarks our proposed method attains state-of-the-art performance in both classification and segmentation CTTA tasks.
****
The STVchrono Dataset: Towards Continuous Change Recognition in Time
Yanjun Sun, Yue Qiu, Mariia Khan, Fumiya Matsuzawa, Kenji Iwata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14111-14120
Recognizing continuous changes offers valuable insights into past historical events supports current trend analysis and facilitates future planning. This knowledge is crucial for a variety of fields such as meteorology and agriculture envir

onmental science urban planning and construction tourism and cultural preservation. Currently available datasets in the field of scene change understanding primarily concentrate on two main tasks: the detection of changed regions within a scene and the linguistic description of the change content. Existing datasets focus on recognizing discrete changes such as adding or deleting an object from two images and largely rely on artificially generated images. Consequently the existing change understanding methods primarily focus on identifying distinct object differences overlooking the importance of continuous gradual changes occurring over extended time intervals. To address the above issues we propose a novel benchmark dataset STVchrono targeting the localization and description of long-term continuous changes in real-world scenes. The dataset consists of 71900 photographs from Google Street View API taken over an 18-year span across 50 cities all over the world. Our STVchrono dataset is designed to support real-world continuous change recognition and description in both image pairs and extended image sequences while also enabling the segmentation of changed regions. We conduct experiments to evaluate state-of-the-art methods on continuous change description and segmentation as well as multimodal Large Language Models for describing changes. Our findings reveal that even the most advanced methods lag human performance emphasizing the need to adapt them to continuously changing real-world scenarios. We hope that our benchmark dataset will further facilitate the research of temporal change recognition in a dynamic world. The STVchrono dataset is available at STVchrono Dataset.
****
SocialCircle: Learning the Angle-based Social Interaction Representation for Pedestrian Trajectory Prediction
Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, Xinge You; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19005-19015
Analyzing and forecasting trajectories of agents like pedestrians and cars in complex scenes has become more and more significant in many intelligent systems and applications. The diversity and uncertainty in socially interactive behaviors among a rich variety of agents make this task more challenging than other deterministic computer vision tasks. Researchers have made a lot of efforts to quantify the effects of these interactions on future trajectories through different mathematical models and network structures but this problem has not been well solved. Inspired by marine animals that localize the positions of their companions underwater through echoes we build a new anglebased trainable social interaction representation named SocialCircle for continuously reflecting the context of social interactions at different angular orientations relative to the target agent. We validate the effect of the proposed SocialCircle by training it along with several newly released trajectory prediction models and experiments show that the SocialCircle not only quantitatively improves the prediction performance but also qualitatively helps better simulate social interactions when forecasting pedestrian trajectories in a way that is consistent with human intuitions.
****
Boosting Neural Representations for Videos with a Conditional Decoder
Xinjie Zhang, Ren Yang, Dailan He, Xingtong Ge, Tongda Xu, Yan Wang, Hongwei Qin, Jun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2556-2566
Implicit neural representations (INRs) have emerged as a promising approach for video storage and processing showing remarkable versatility across various video tasks. However existing methods often fail to fully leverage their representation capabilities primarily due to inadequate alignment of intermediate features during target frame decoding. This paper introduces a universal boosting framework for current implicit video representation approaches. Specifically we utilize a conditional decoder with a temporal-aware affine transform module which uses the frame index as a prior condition to effectively align intermediate features with target frames. Besides we introduce a sinusoidal NeRV-like block to generate diverse intermediate features and achieve a more balanced parameter distribution thereby enhancing the model's capacity. With a high-frequency information-pres

erving reconstruction loss our approach successfully boosts multiple baseline INRs in the reconstruction quality and convergence speed for video regression and exhibits superior inpainting and interpolation results. Further we integrate a consistent entropy minimization technique and develop video codecs based on these boosted INRs. Experiments on the UVG dataset confirm that our enhanced codecs significantly outperform baseline INRs and offer competitive rate-distortion performance compared to traditional and learning-based codecs. Code is available at https://github.com/Xinjie-Q/Boosting-NeRV.
****
Dual-Enhanced Coreset Selection with Class-wise Collaboration for Online Blurry Class Incremental Learning
Yutian Luo, Shiqi Zhao, Haoran Wu, Zhiwu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23995-24004
Traditional online class incremental learning assumes class sets in different tasks are disjoint. However recent works have shifted towards a more realistic scenario where tasks have shared classes creating blurred task boundaries. Under this setting although existing approaches could be directly applied challenges like data imbalance and varying class-wise data volumes complicate the critical coreset selection used for replay. To tackle these challenges we introduce DECO (Dual-Enhanced Coreset Selection with Class-wise Collaboration) an approach that starts by establishing a class-wise balanced memory to address data imbalances followed by a tailored class-wise gradient-based similarity scoring system for refined coreset selection strategies with reasonable score guidance to all classes. DECO is distinguished by two main strategies: (1) Collaborative Diverse Score Guidance that mitigates biased knowledge in less-exposed classes through guidance from well-established classes simultaneously consolidating the knowledge in the established classes to enhance overall stability. (2) Adaptive Similarity Score Constraint that relaxes constraints between class types boosting learning plasticity for less-exposed classes and assisting well-established classes in defining clearer boundaries thereby improving overall plasticity. Overall DECO helps effectively identify critical coreset samples improving learning stability and plasticity across all classes. Extensive experiments are conducted on four benchmark datasets to demonstrate the effectiveness and superiority of DECO over other competitors under this online blurry class incremental learning setting.
****
From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations
Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, Alexander Richard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1001-1010
We present a framework for generating full-bodied photorealistic avatars that gesture according to the conversational dynamics of a dyadic interaction. Given speech audio we output multiple possibilities of gestural motion for an individual including face body and hands. The key behind our method is in combining the benefits of sample diversity from vector quantization with the high-frequency details obtained through diffusion to generate more dynamic expressive motion. We visualize the generated motion using highly photorealistic avatars that can express crucial nuances in gestures (e.g. sneers and smirks). To facilitate this line of research we introduce a first-of-its-kind multi-view conversational dataset that allows for photorealistic reconstruction. Experiments show our model generates appropriate and diverse gestures outperforming both diffusion- and VQ-only methods. Furthermore our perceptual evaluation highlights the importance of photorealism (vs. meshes) in accurately assessing subtle motion details in conversational gestures. Code and dataset available on project page.
****
Single-View Scene Point Cloud Human Grasp Generation
Yan-Kang Wang, Chengyi Xing, Yi-Lin Wei, Xiao-Ming Wu, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 831-841
In this work we explore a novel task of generating human grasps based on single-view scene point clouds which more accurately mirrors the typical real-world sit

uation of observing objects from a single viewpoint. Due to the incompleteness o
f object point clouds and the presence of numerous scene points the generated ha
nd is prone to penetrating into the invisible parts of the object and the model
is easily affected by scene points. Thus we introduce S2HGrasp a framework compo
sed of two key modules: the Global Perception module that globally perceives par
tial object point clouds and the DiffuGrasp module designed to generate high-qua
lity human grasps based on complex inputs that include scene points. Additionall
y we introduce S2HGD dataset which comprises approximately 99000 single-object s
ingle-view scene point clouds of 1668 unique objects each annotated with one hum
an grasp. Our extensive experiments demonstrate that S2HGrasp can not only gener
ate natural human grasps regardless of scene points but also effectively prevent
 penetration between the hand and invisible parts of the object. Moreover our mo
del showcases strong generalization capability when applied to unseen objects. O
ur code and dataset are available at https://github.com/iSEE-Laboratory/S2HGrasp
.
****
One-step Diffusion with Distribution Matching Distillation
Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William
 T. Freeman, Taesung Park; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2024, pp. 6613-6623
Diffusion models generate high-quality images but require dozens of forward pass
es. We introduce Distribution Matching Distillation (DMD) a procedure to transfo
rm a diffusion model into a one-step image generator with minimal impact on imag
e quality. We enforce the one-step image generator match the diffusion model at
distribution level by minimizing an approximate KL divergence whose gradient can
 be expressed as the difference between 2 score functions one of the target dist
ribution and the other of the synthetic distribution being produced by our one-s
tep generator. The score functions are parameterized as two diffusion models tra
ined separately on each distribution. Combined with a simple regression loss mat
ching the large-scale structure of the multi-step diffusion outputs our method o
utperforms all published few-step diffusion approaches reaching 2.62 FID on Imag
eNet 64x64 and 11.49 FID on zero-shot COCO-30k comparable to Stable Diffusion bu
t orders of magnitude faster. Utilizing FP16 inference our model can generate im
ages at 20 FPS on modern hardware.
****
Cyclic Learning for Binaural Audio Generation and Localization
Zhaojian Li, Bin Zhao, Yuan Yuan; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2024, pp. 26669-26678
Binaural audio is obtained by simulating the biological structure of human ears
which plays an important role in artificial immersive spaces. A promising approa
ch is to utilize mono audio and corresponding vision to synthesize binaural audi
o thereby avoiding expensive binaural audio recording. However most existing met
hods directly use the entire scene as a guide ignoring the correspondence betwee
n sounds and sounding objects. In this paper we advocate generating binaural aud
io using fine-grained raw waveform and object-level visual information as guidan
ce. Specifically we propose a Cyclic Locating-and-UPmixing (CLUP) framework that
 jointly learns visual sounding object localization and binaural audio generatio
n. Visual sounding object localization establishes the correspondence between sp
ecific visual objects and sound modalities which provides object-aware guidance
to improve binaural generation performance. Meanwhile the spatial information co
ntained in the generated binaural audio can further improve the performance of s
ounding object localization. In this case visual sounding object localization an
d binaural audio generation can achieve cyclic learning and benefit from each ot
her. Experimental results demonstrate that on the FAIR-Play benchmark dataset ou
r method is significantly ahead of the existing baselines in multiple evaluation
 metrics (STFT\downarrow: 0.787 vs. 0.851 ENV\downarrow: 0.128 vs. 0.134 WAV\dow
narrow: 5.244 vs. 5.684 SNR\uparrow: 7.546 vs. 7.044).
****
Neighbor Relations Matter in Video Scene Detection
Jiawei Tan, Hongxing Wang, Jiaxin Li, Zhilong Ou, Zhangbin Qian; Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 18473-18482
Video scene detection aims to temporally link shots for obtaining semantically c
ompact scenes. It is essential for this task to capture scene-distinguishable af
finity among shots by similarity assessment. However most methods relies on ordi
nary shot-to-shot similarities which may inveigle similar shots into being linke
d even though they are from different scenes and meanwhile hinder dissimilar sho
ts from being blended into a complete scene. In this paper we propose NeighborNe
t to inject shot contexts into shot-to-shot similarities through carefully explo
ring the relations between semantic/temporal neighbors of shots over a local tim
e period. In this way shot-to-shot similarities are remeasured as semantic/tempo
ral neighbor-aware similarities so that NeighborNet can learn context embedding
into shot features using graph convolutional network. As a result not only do th
e learned shot features suppress the affinity among similar shots from different
 scenes but they also promote the affinity among dissimilar shots in the same sc
ene. Experimental results on public benchmark datasets show that our proposed Ne
ighborNet yields substantial improvements in video scene detection especially ou
tperforms released state-of-the-arts by at least 6% in Average Precision (AP). T
he code is available at https://github.com/ExMorgan-Alter/NeighborNet.
****
Rethinking Human Motion Prediction with Symplectic Integral
Haipeng Chen, Kedi Lyu, Zhenguang Liu, Yifang Yin, Xun Yang, Yingda Lyu; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 2134-2143
Long-term and accurate forecasting is the long-standing pursuit of the human mot
ion prediction task. Existing methods typically suffer from dramatic degradation
 in prediction accuracy with the increasing prediction horizon. It comes down to
 two reasons:1? Insufficient numerical stability.Unforeseen high noise and compl
ex feature relationships in the data. 2? Inadequate modeling stability. Unreason
able step sizes and undesirable parameter updates in the prediction.In this pape
r we design a novel and symplectic integral-inspired framework named symplectic
integral neural network (SINN) which engages symplectic trajectories to optimize
 the pose representation and employs a stable symplectic operator to alternately
 model the dynamic context. Specifically we design a Symplectic Representation E
ncoder that performs on enhanced human pose representation to obtain trajectorie
s on the symplectic manifold ensuring numerical stability based on Hamiltonian m
echanics and symplectic spatial splitting algorithm. We further present the Symp
lectic Temporal Aggregation module in the light of the symplectic temporal split
ting algorithm which splits the long-term prediction into multiple accurate shor
t-term predictions generated by a symplectic operator to secure modeling stabili
ty. Moreover our approach is model-agnostic and can be efficiently integrated wi
th different physical dynamics models.The experimental results demonstrate that
our method achieves the new state-of-the-art outperforming existing methods by l
arge margins:20.1%on Human3.6M16.7%on CUM Mocap and 10.2% on 3DPW.
****
Text-to-Image Diffusion Models are Great Sketch-Photo Matchmakers
Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xi
ang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2024, pp. 16826-16837
This paper for the first time explores text-to-image diffusion models for Zero-S
hot Sketch-based Image Retrieval (ZS-SBIR). We highlight a pivotal discovery: th
e capacity of text-to-image diffusion models to seamlessly bridge the gap betwee
n sketches and photos. This proficiency is underpinned by their robust cross-mod
al capabilities and shape bias findings that are substantiated through our pilot
 studies. In order to harness pre-trained diffusion models effectively we introd
uce a straightforward yet powerful strategy focused on two key aspects: selectin
g optimal feature layers and utilising visual and textual prompts. For the forme
r we identify which layers are most enriched with information and are best suite
d for the specific retrieval requirements (category-level or fine-grained). Then
 we employ visual and textual prompts to guide the model's feature extraction pr

ocess enabling it to generate more discriminative and contextually relevant cross-modal representations. Extensive experiments on several benchmark datasets validate significant performance improvements.
****
Mudslide: A Universal Nuclear Instance Segmentation Method
Jun Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11673-11682
Nuclear instance segmentation has played a critical role in pathology image analysis. The main challenges arise from the difficulty in accurately segmenting densely overlapping instances and the high cost of precise mask-level annotations. Existing fully-supervised nuclear instance segmentation methods such as boundary-based methods struggle to capture differences between overlapping instances and thus fail in densely distributed blurry regions. They also face challenges transitioning to point supervision where annotations are simple and effective. Inspired by natural mudslides we propose a universal method called Mudslide that uses simple representations to characterize differences between different instances and can easily be extended from fully-supervised to point-supervised. oncretely we introduce a collapse field and leverage it to construct a force map and initial boundary enabling a distinctive representation for each instance. Each pixel is assigned a collapse force with distinct directions between adjacent instances. Starting from the initial boundary Mudslide executes a pixel-by-pixel collapse along various force directions. Pixels that collapse into the same region are considered as one instance concurrently accounting for both inter-instance distinctions and intra-instance coherence. Experiments on public datasets show superior performance in both fully-supervised and point-supervised tasks.
****
CPGA: Coding Priors-Guided Aggregation Network for Compressed Video Quality Enhancement
Qiang Zhu, Jinhua Hao, Yukang Ding, Yu Liu, Qiao Mo, Ming Sun, Chao Zhou, Shuyuan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2964-2974
Recently numerous approaches have achieved notable success in compressed video quality enhancement (VQE). However these methods usually ignore the utilization of valuable coding priors inherently embedded in compressed videos such as motion vectors and residual frames which carry abundant temporal and spatial information. To remedy this problem we propose the Coding Priors-Guided Aggregation (CPGA) network to utilize temporal and spatial information from coding priors. The CPGA mainly consists of an inter-frame temporal aggregation (ITA) module and a multi-scale non-local aggregation (MNA) module. Specifically the ITA module aggregates temporal information from consecutive frames and coding priors while the MNA module globally captures spatial information guided by residual frames. In addition to facilitate research in VQE task we newly construct the Video Coding Priors (VCP) dataset comprising 300 videos with various coding priors extracted from corresponding bitstreams. It remedies the shortage of previous datasets on the lack of coding information. Experimental results demonstrate the superiority of our method compared to existing state-of-the-art methods. The code and dataset will be released at https://github.com/VQE-CPGA/CPGA.
****
MicroCinema: A Divide-and-Conquer Approach for Text-to-Video Generation
Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, Kai Qiu, Yuhui Yuan, Xiaoyan Sun, Chong Luo, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8414-8424
We present MicroCinema a straightforward yet effective framework for high-quality and coherent text-to-video generation. Unlike existing approaches that align text prompts with video directly MicroCinema introduces a Divide-and-Conquer strategy which divides the text-to-video into a two-stage process: text-to-image generation and image&text-to-video generation. This strategy offers two significant advantages. a) It allows us to take full advantage of the recent advances in text-to-image models such as Stable Diffusion Midjourney and DALLE to generate pho

torealistic and highly detailed images. b) Leveraging the generated image the mo
del can allocate less focus to fine-grained appearance details prioritizing the
efficient learning of motion dynamics. To implement this strategy effectively we
 introduce two core designs. First we propose the Appearance Injection Network e
nhancing the preservation of the appearance of the given image. Second we introd
uce the Appearance Noise Prior a novel mechanism aimed at maintaining the capabi
lities of pre-trained 2D diffusion models. These design elements empower MicroCi
nema to generate high-quality videos with precise motion guided by the provided
text prompts. Extensive experiments demonstrate the superiority of the proposed
framework. Concretely MicroCinema achieves SOTA zero-shot FVD of 342.86 on UCF-1
01 and 377.40 on MSR-VTT.
****
Learning Instance-Aware Correspondences for Robust Multi-Instance Point Cloud Re
gistration in Cluttered Scenes
Zhiyuan Yu, Zheng Qin, Lintao Zheng, Kai Xu; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19605-19614
Multi-instance point cloud registration estimates the poses of multiple instance
s of a model point cloud in a scene point cloud. Extracting accurate point corre
spondences is to the center of the problem. Existing approaches usually treat th
e scene point cloud as a whole overlooking the separation of instances. Therefor
e point features could be easily polluted by other points from the background or
 different instances leading to inaccurate correspondences oblivious to separate
 instances especially in cluttered scenes. In this work we propose MIRETR Multi-
Instance REgistration TRansformer a coarse-to-fine approach to the extraction of
 instance-aware correspondences. At the coarse level it jointly learns instance-
aware superpoint features and predicts per-instance masks. With instance masks t
he influence from outside of the instance being concerned is minimized such that
 highly reliable superpoint correspondences can be extracted. The superpoint cor
respondences are then extended to instance candidates at the fine level accordin
g to the instance masks. At last an efficient candidate selection and refinement
 algorithm is devised to obtain the final registrations. Extensive experiments o
n three public benchmarks demonstrate the efficacy of our approach. In particula
r MIRETR outperforms the state of the arts by 16.6 points on F1 score on the cha
llenging ROBI benchmark. Code and models are available at https://github.com/zhi
yuanYU134/MIRETR
****
Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Ima
ge Inpainting
Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, Yong Rui; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8038
-8047
Denoising diffusion probabilistic models (DDPMs) for image inpainting aim to add
 the noise to the texture of the image during the forward process and recover th
e masked regions with the unmasked ones of the texture via the reverse denoising
 process. Despite the meaningful semantics generation the existing arts suffer f
rom the semantic discrepancy between the masked and unmasked regions since the s
emantically dense unmasked texture fails to be completely degraded while the mas
ked regions turn to the pure noise in diffusion process leading to the large dis
crepancy between them. In this paper we aim to answer how the unmasked semantics
 guide the texture denoising process; together with how to tackle the semantic d
iscrepancy to facilitate the consistent and meaningful semantics generation. To
this end we propose a novel structure-guided diffusion model for image inpaintin
g named StrDiffusion to reformulate the conventional texture denoising process u
nder the structure guidance to derive a simplified denoising objective for image
 inpainting while revealing: 1) the semantically sparse structure is beneficial
to tackle the semantic discrepancy in the early stage while the dense texture ge
nerates the reasonable semantics in the late stage; 2) the semantics from the un
masked regions essentially offer the time-dependent structure guidance for the t
exture denoising process benefiting from the time-dependent sparsity of the stru
cture semantics. For the denoising process a structure-guided neural network is

trained to estimate the simplified denoising objective by exploiting the consist
ency of the denoised structure between masked and unmasked regions. Besides we d
evise an adaptive resampling strategy as a formal criterion as whether the struc
ture is competent to guide the texture denoising process while regulate their se
mantic correlations. Extensive experiments validate the merits of StrDiffusion o
ver the state-of-the-arts. Our code is available at https://github.com/htyjers/S
trDiffusion.
****
Modeling Multimodal Social Interactions: New Challenges and Baselines with Dense
ly Aligned Representations
Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, James M. Rehg; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 14585-14595
Understanding social interactions involving both verbal and non-verbal cues is e
ssential for effectively interpreting social situations. However most prior work
s on multimodal social cues focus predominantly on single-person behaviors or re
ly on holistic visual representations that are not aligned to utterances in mult
i-party environments. Consequently they are limited in modeling the intricate dy
namics of multi-party interactions. In this paper we introduce three new challen
ging tasks to model the fine-grained dynamics between multiple people: speaking
target identification pronoun coreference resolution and mentioned player predic
tion. We contribute extensive data annotations to curate these new challenges in
 social deduction game settings. Furthermore we propose a novel multimodal basel
ine that leverages densely aligned language-visual representations by synchroniz
ing visual features with their corresponding utterances. This facilitates concur
rently capturing verbal and non-verbal cues pertinent to social reasoning. Exper
iments demonstrate the effectiveness of the proposed approach with densely align
ed multimodal representations in modeling fine-grained social interactions. Proj
ect website: https://sangmin-git.github.io/projects/MMSI.
****
COCONut: Modernizing COCO Segmentation
Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, Liang-Chieh Chen; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 21863-21873
In recent decades the vision community has witnessed remarkable progress in visu
al recognition partially owing to advancements in dataset benchmarks. Notably th
e established COCO benchmark has propelled the development of modern detection a
nd segmentation systems. However the COCO segmentation benchmark has seen compar
atively slow improvement over the last decade. Originally equipped with coarse p
olygon annotations for thing instances it gradually incorporated coarse superpix
el annotations for stuff regions which were subsequently heuristically amalgamat
ed to yield panoptic segmentation annotations. These annotations executed by dif
ferent groups of raters have resulted not only in coarse segmentation masks but
also in inconsistencies between segmentation types. In this study we undertake a
 comprehensive reevaluation of the COCO segmentation annotations. By enhancing t
he annotation quality and expanding the dataset to encompass 383K images with mo
re than 5.18M panoptic masks we introduce COCONut the COCO Next Universal segmen
Tation dataset. COCONut harmonizes segmentation annotations across semantic inst
ance and panoptic segmentation with meticulously crafted high-quality masks and
establishes a robust benchmark for all segmentation tasks. To our knowledge COCO
Nut stands as the inaugural large-scale universal segmentation dataset verified
by human raters. We anticipate that the release of COCONut will significantly co
ntribute to the community's ability to assess the progress of novel neural netwo
rks.
****
Semantic Line Combination Detector
Jinwon Ko, Dongkwon Jin, Chang-Su Kim; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28066-28075
A novel algorithm called semantic line combination detector (SLCD) to find an op
timal combination of semantic lines is proposed in this paper. It processes all

lines in each line combination at once to assess the overall harmony of the line
s. First we generate various line combinations from reliable lines. Second we es
timate the score of each line combination and determine the best one. Experiment
al results demonstrate that the proposed SLCD outperforms existing semantic line
 detectors on various datasets. Moreover it is shown that SLCD can be applied ef
fectively to three vision tasks of vanishing point detection symmetry axis detec
tion and composition-based image retrieval. Our codes are available at https://g
ithub.com/Jinwon-Ko/SLCD.
****
Prompt-Driven Dynamic Object-Centric Learning for Single Domain Generalization
Deng Li, Aming Wu, Yaowei Wang, Yahong Han; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17606-17615
Single-domain generalization aims to learn a model from single source domain dat
a attaining generalized performance on other unseen target domains. Existing wor
ks primarily focus on improving the generalization ability of static networks. H
owever static networks are unable to dynamically adapt to the diverse variations
 in different image scenes leading to limited generalization capability. Differe
nt scenes exhibit varying levels of complexity and the complexity of images furt
her varies significantly in cross-domain scenarios. In this paper we propose a d
ynamic object-centric perception network based on prompt learning aiming to adap
t to the variations in image complexity. Specifically we propose an object-centr
ic gating module based on prompt learning to focus attention on the object-centr
ic features guided by the various scene prompts. Then with the object-centric ga
ting masks the dynamic selective module dynamically selects highly correlated fe
ature regions in both spatial and channel dimensions enabling the model to adapt
ively perceive object-centric relevant features thereby enhancing the generaliza
tion capability. Extensive experiments were conducted on single-domain generaliz
ation tasks in image classification and object detection. The experimental resul
ts demonstrate that our approach outperforms state-of-the-art methods which vali
dates the effectiveness and versatility of our proposed method.
****
Dual Pose-invariant Embeddings: Learning Category and Object-specific Discrimina
tive Representations for Recognition and Retrieval
Rohan Sarkar, Avinash Kak; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2024, pp. 17077-17085
In the context of pose-invariant object recognition and retrieval we demonstrate
 that it is possible to achieve significant improvements in performance if both
the category-based and the object-identity-based embeddings are learned simultan
eously during training. In hindsight that sounds intuitive because learning abou
t the categories is more fundamental than learning about the individual objects
that correspond to those categories. However to the best of what we know no prio
r work in pose invariant learning has demonstrated this effect. This paper prese
nts an attention-based dual-encoder architecture with specially designed loss fu
nctions that optimize the inter- and intra-class distances simultaneously in two
 different embedding spaces one for the category embeddings and the other for th
e object level embeddings. The loss functions we have proposed are pose-invarian
t ranking losses that are designed to minimize the intra-class distances and max
imize the inter-class distances in the dual representation spaces. We demonstrat
e the power of our approach with three challenging multi-view datasets ModelNet-
40 ObjectPI and FG3D. With our dual approach for single view object recognition
we outperform the previous best by 20.0% on ModelNet40 2.0% on ObjectPI and 46.5
% on FG3D. On the other hand for single-view object retrieval we outperform the
previous best by 33.7% on ModelNet40 18.8% on ObjectPI and 56.9% on FG3D.
****
vid-TLDR: Training Free Token Merging for Light-weight Video Transformer
Joonmyung Choi, Sanghyeok Lee, Jaewon Chu, Minhyuk Choi, Hyunwoo J. Kim; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 18771-18781
Video Transformers have become the prevalent solution for various video downstre
am tasks with superior expressive power and flexibility. However these video tra

nsformers suffer from heavy computational costs induced by the massive number of tokens across the entire video frames which has been the major barrier to training the model. Further the patches irrelevant to the main contents e.g. backgrounds degrade the generalization performance of models. To tackle these issues we propose training free token merging for lightweight video Transformer (vid-TLDR) that aims to enhance the efficiency of video Transformers by merging the background tokens without additional training. For vid-TLDR we introduce a novel approach to capture the salient regions in videos only with the attention map. Further we introduce the saliency-aware token merging strategy by dropping the background tokens and sharpening the object scores. Our experiments show that vid-TLDR significantly mitigates the computational complexity of video Transformers while achieving competitive performance compared to the base model without vid-TLDR. Code is available at https://github.com/mlvlab/vid-TLDR.
****
DRESS: Instructing Large Vision-Language Models to Align and Interact with Humans via Natural Language Feedback
Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, Ajay Divakaran; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14239-14250
We present DRESS a large vision language model (LVLM) that innovatively exploits Natural Language feedback (NLF) from Large Language Models to enhance its alignment and interactions by addressing two key limitations in the state-of-the-art LVLMs. First prior LVLMs generally rely only on the instruction finetuning stage to enhance alignment with human preferences. Without incorporating extra feedback they are still prone to generate unhelpful hallucinated or harmful responses. Second while the visual instruction tuning data is generally structured in a multi-turn dialogue format the connections and dependencies among consecutive conversational turns are weak. This reduces the capacity for effective multi-turn interactions. To tackle these we propose a novel categorization of the NLF into two key types: critique and refinement. The critique NLF identifies the strengths and weaknesses of the responses and is used to align the LVLMs with human preferences. The refinement NLF offers concrete suggestions for improvement and is adopted to improve the interaction ability of the LVLMs-- which focuses on LVLMs' ability to refine responses by incorporating feedback in multi-turn interactions. To address the non-differentiable nature of NLF we generalize conditional reinforcement learning for training. Our experimental results demonstrate that DRESS can generate more helpful (9.76%) honest (11.52%) and harmless (21.03%) responses and more effectively learn from feedback during multi-turn interactions compared to SOTA LVLMs.
****
Makeup Prior Models for 3D Facial Makeup Estimation and Applications
Xingchao Yang, Takafumi Taketomi, Yuki Endo, Yoshihiro Kanamori; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2165-2176
In this work we introduce two types of makeup prior models to extend existing 3D face prior models: PCA-based and StyleGAN2-based priors. The PCA-based prior model is a linear model that is easy to construct and is computationally efficient. However it retains only low-frequency information. Conversely the StyleGAN2-based model can represent high-frequency information with relatively higher computational cost than the PCA-based model. Although there is a trade-off between the two models both are applicable to 3D facial makeup estimation and related applications. By leveraging makeup prior models and designing a makeup consistency module we effectively address the challenges that previous methods faced in robustly estimating makeup particularly in the context of handling self-occluded faces. In experiments we demonstrate that our approach reduces computational costs by several orders of magnitude achieving speeds up to 180 times faster. In addition by improving the accuracy of the estimated makeup we confirm that our methods are highly advantageous for various 3D facial makeup applications such as 3D makeup face reconstruction user-friendly makeup editing makeup transfer and interpolation.

****
Salience DETR: Enhancing Detection Transformer with Hierarchical Salience Filtering Refinement
Xiuquan Hou, Meiqin Liu, Senlin Zhang, Ping Wei, Badong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17574-17583
DETR-like methods have significantly increased detection performance in an end-to-end manner. The mainstream two-stage frameworks of them perform dense self-attention and select a fraction of queries for sparse cross-attention which is proven effective for improving performance but also introduces a heavy computational burden and high dependence on stable query selection. This paper demonstrates that suboptimal two-stage selection strategies result in scale bias and redundancy due to the mismatch between selected queries and objects in two-stage initialization. To address these issues we propose hierarchical salience filtering refinement which performs transformer encoding only on filtered discriminative queries for a better trade-off between computational efficiency and precision. The filtering process overcomes scale bias through a novel scale-independent salience supervision. To compensate for the semantic misalignment among queries we introduce elaborate query refinement modules for stable two-stage initialization. Based on above improvements the proposed Salience DETR achieves significant improvements of +4.0% AP +0.2% AP +4.4% AP on three challenging task-specific detection datasets as well as 49.2% AP on COCO 2017 with less FLOPs. The code is available at https://github.com/xiuqhou/Salience-DETR.
****
Towards More Unified In-context Visual Understanding
Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Tao Gong, Bin Liu, Shengwei Xu, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13362-13372
The rapid advancement of large language models (LLMs) has accelerated the emergence of in-context learning (ICL) as a cutting-edge approach in the natural language processing domain. Recently ICL has been employed in visual understanding tasks such as semantic segmentation and image captioning yielding promising results. However existing visual ICL framework can not enable producing content across multiple modalities which limits their potential usage scenarios. To address this issue we present a new ICL framework for visual understanding with multi-modal output enabled. First we quantize and embed both text and visual prompt into a unified representational space structured as interleaved in-context sequences. Then a decoder-only sparse transformer architecture is employed to perform generative modeling on them facilitating in-context learning. Thanks to this design the model is capable of handling in-context vision understanding tasks with multimodal output in a unified pipeline. Experimental results demonstrate that our model achieves competitive performance compared with specialized models and previous ICL baselines. Overall our research takes a further step toward unified multimodal in-context learning.
****
F3Loc: Fusion and Filtering for Floorplan Localization
Changan Chen, Rui Wang, Christoph Vogel, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18029-18038
In this paper we propose an efficient data-driven solution to self-localization within a floorplan. Floorplan data is readily available long-term persistent and inherently robust to changes in the visual appearance. Our method does not require retraining per map and location or demand a large database of images of the area of interest. We propose a novel probabilistic model consisting of an observation and a novel temporal filtering module. Operating internally with an efficient ray-based representation the observation module consists of a single and a multiview module to predict horizontal depth from images and fuses their results to benefit from advantages offered by either methodology. Our method operates on conventional consumer hardware and overcomes a common limitation of competing methods that often demand upright images. Our full system meets real-time require

ments while outperforming the state-of-the-art by a significant margin.
****
ReconFusion: 3D Reconstruction with Diffusion Priors
Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, Aleksander Ho?y?ski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21551-21561
3D reconstruction methods such as Neural Radiance Fields (NeRFs) excel at rendering photorealistic novel views of complex scenes. However recovering a high-quality NeRF typically requires tens to hundreds of input images resulting in a time-consuming capture process. We present ReconFusion to reconstruct real-world scenes using only a few photos. Our approach leverages a diffusion prior for novel view synthesis trained on synthetic and multiview datasets which regularizes a NeRF-based 3D reconstruction pipeline at novel camera poses beyond those captured by the set of input images. Our method synthesizes realistic geometry and texture in underconstrained regions while preserving the appearance of observed regions. We perform an extensive evaluation across various real-world datasets including forward-facing and 360-degree scenes demonstrating significant performance improvements over previous few-view NeRF reconstruction approaches. Please see our project page at reconfusion.github.io.
****
I'M HOI: Inertia-aware Monocular Capture of 3D Human-Object Interactions
Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, Lan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 729-741
We are living in a world surrounded by diverse and "smart" devices with rich modalities of sensing ability. Conveniently capturing the interactions between us humans and these objects remains far-reaching. In this paper we present I'm-HOI a monocular scheme to faithfully capture the 3D motions of both the human and object in a novel setting: using a minimal amount of RGB camera and object-mounted Inertial Measurement Unit (IMU). It combines general motion inference and category-aware refinement. For the former we introduce a holistic human-object tracking method to fuse the IMU signals and the RGB stream and progressively recover the human motions and subsequently the companion object motions. For the latter we tailor a category-aware motion diffusion model which is conditioned on both the raw IMU observations and the results from the previous stage under over-parameterization representation. It significantly refines the initial results and generates vivid body hand and object motions. Moreover we contribute a large dataset with ground truth human and object motions dense RGB inputs and rich object-mounted IMU measurements. Extensive experiments demonstrate the effectiveness of I'm-HOI under a hybrid capture setting. Our dataset and code will be released to the community.
****
Dynamic Policy-Driven Adaptive Multi-Instance Learning for Whole Slide Image Classification
Tingting Zheng, Kui Jiang, Hongxun Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8028-8037
Multi-Instance Learning (MIL) has shown impressive performance for histopathology whole slide image (WSI) analysis using bags or pseudo-bags. It involves instance sampling feature representation and decision-making. However existing MIL-based technologies at least suffer from one or more of the following problems: 1) requiring high storage and intensive pre-processing for numerous instances (sampling); 2) potential over-fitting with limited knowledge to predict bag labels (feature representation); 3) pseudo-bag counts and prior biases affect model robustness and generalizability (decision-making). Inspired by clinical diagnostics using the past sampling instances can facilitate the final WSI analysis but it is barely explored in prior technologies. To break free these limitations we integrate the dynamic instance sampling and reinforcement learning into a unified framework to improve the instance selection and feature aggregation forming a novel Dynamic Policy Instance Selection (DPIS) scheme for better and more credible dec

ision-making. Specifically the measurement of feature distance and reward functi
on are employed to boost continuous instance sampling. To alleviate the over-fit
ting we explore the latent global relations among instances for more robust and
discriminative feature representation while establishing reward and punishment m
echanisms to correct biases in pseudo-bags using contrastive learning. These str
ategies form the final Dynamic Policy-Driven Adaptive Multi-Instance Learning (P
AMIL) method for WSI tasks. Extensive experiments reveal that our PAMIL method o
utperforms the state-of-the-art by 3.8% on CAMELYON16 and 4.4% on TCGA lung canc
er datasets.
****
InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Li
nguistic Tasks
Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Q
inglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, Jifeng
Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2024, pp. 24185-24198
The exponential growth of large language models (LLMs) has opened up numerous po
ssibilities for multi-modal AGI systems. However the progress in vision and visi
on-language foundation models which are also critical elements of multi-modal AG
I has not kept pace with LLMs. In this work we design a large-scale vision-langu
age foundation model (InternVL) which scales up the vision foundation model to 6
 billion parameters and progressively aligns it with the LLM using web-scale ima
ge-text data from various sources. This model can be broadly applied to and achi
eve state-of-the-art performance on 32 generic visual-linguistic benchmarks incl
uding visual perception tasks such as image-level or pixel-level recognition vis
ion-language tasks such as zero-shot image/video classification zero-shot image/
video-text retrieval and link with LLMs to create multi-modal dialogue systems.
It has powerful visual capabilities and can be a good alternative to the ViT-22B
. We hope that our research could contribute to the development of multi-modal l
arge models.
****
Multi-View Attentive Contextualization for Multi-View 3D Object Detection
Xianpeng Liu, Ce Zheng, Ming Qian, Nan Xue, Chen Chen, Zhebin Zhang, Chen Li, Ti
anfu Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2024, pp. 16688-16698
We present Multi-View Attentive Contextualization (MvACon) a simple yet effectiv
e method for improving 2D-to-3D feature lifting in query-based multi-view 3D (MV
3D) object detection. Despite remarkable progress witnessed in the field of quer
y-based MV3D object detection prior art often suffers from either the lack of ex
ploiting high-resolution 2D features in dense attention-based lifting due to hig
h computational costs or from insufficiently dense grounding of 3D queries to mu
lti-scale 2D features in sparse attention-based lifting. Our proposed MvACon hit
s the two birds with one stone using a representationally dense yet computationa
lly sparse attentive feature contextualization scheme that is agnostic to specif
ic 2D-to-3D feature lifting approaches. In experiments the proposed MvACon is th
oroughly tested on the nuScenes benchmark using both the BEVFormer and its recen
t 3D deformable attention (DFA3D) variant as well as the PETR showing consistent
 detection performance improvement especially in enhancing performance in locati
on orientation and velocity prediction. It is also tested on the Waymo-mini benc
hmark using BEVFormer with similar improvement. We qualitatively and quantitativ
ely show that global cluster-based contexts effectively encode dense scene-level
 contexts for MV3D object detection. The promising results of our proposed MvACo
n reinforces the adage in computer vision "(contextualized) feature matters".
****
MemSAM: Taming Segment Anything Model for Echocardiography Video Segmentation
Xiaolong Deng, Huisi Wu, Runhao Zeng, Jing Qin; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9622-9631
We propose a novel echocardiographical video segmentation model by adapting SAM
to medical videos to address some long-standing challenges in ultrasound video s
egmentation including (1) massive speckle noise and artifacts (2) extremely ambi

guous boundaries and (3) large variations of targeting objects across frames. Th e core technique of our model is a temporal-aware and noise-resilient prompting scheme. Specifically we employ a space-time memory that contains both spatial an d temporal information to prompt the segmentation of current frame and thus we c all the proposed model as MemSAM. In prompting the memory carrying temporal cues sequentially prompt the video segmentation frame by frame. Meanwhile as the mem ory prompt propagates high-level features it avoids the issue of misidentificati on caused by mask propagation and improves representation consistency. To addres s the challenge of speckle noise we further propose a memory reinforcement mecha nism which leverages predicted masks to improve the quality of the memory before storing it. We extensively evaluate our method on two public datasets and demon strate state-of-the-art performance compared to existing models. Particularly ou r model achieves comparable performance with fully supervised approaches with li mited annotations. Codes are available at https://github.com/dengxl0520/MemSAM.
****
LiDAR4D: Dynamic Neural Fields for Novel Space-time View LiDAR Synthesis
Zehan Zheng, Fan Lu, Weiyi Xue, Guang Chen, Changjun Jiang; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5145-5154
Although neural radiance fields (NeRFs) have achieved triumphs in image novel vi ew synthesis (NVS) LiDAR NVS remains largely unexplored. Previous LiDAR NVS meth ods employ a simple shift from image NVS methods while ignoring the dynamic natu re and the large-scale reconstruction problem of LiDAR point clouds. In light of this we propose LiDAR4D a differentiable LiDAR-only framework for novel space-t ime LiDAR view synthesis. In consideration of the sparsity and large-scale chara cteristics we design a 4D hybrid representation combined with multi-planar and g rid features to achieve effective reconstruction in a coarse-to-fine manner. Fur thermore we introduce geometric constraints derived from point clouds to improve temporal consistency. For the realistic synthesis of LiDAR point clouds we inco rporate the global optimization of ray-drop probability to preserve cross-region patterns. Extensive experiments on KITTI-360 and NuScenes datasets demonstrate the superiority of our method in accomplishing geometry-aware and time-consisten t dynamic reconstruction. Codes are available at https://github.com/ispc-lab/LiD AR4D.
****
Exploiting Diffusion Prior for Generalizable Dense Prediction
Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp . 7861-7871
Contents generated by recent advanced Text-to-Image (T2I) diffusion models are s ometimes too imaginative for existing off-the-shelf dense predictors to estimate due to the immitigable domain gap. We introduce DMP a pipeline utilizing pre-tr ained T2I models as a prior for dense prediction tasks. To address the misalignm ent between deterministic prediction tasks and stochastic T2I models we reformul ate the diffusion process through a sequence of interpolations establishing a de terministic mapping between input RGB images and output prediction distributions . To preserve generalizability we use low-rank adaptation to fine-tune pre-train ed models. Extensive experiments across five tasks including 3D property estimat ion semantic segmentation and intrinsic image decomposition showcase the efficac y of the proposed method. Despite limited-domain training data the approach yiel ds faithful estimations for arbitrary images surpassing existing state-of-the-ar t algorithms.
****
PI3D: Efficient Text-to-3D Generation with Pseudo-Image Diffusion
Ying-Tian Liu, Yuan-Chen Guo, Guan Luo, Heyi Sun, Wei Yin, Song-Hai Zhang; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2024, pp. 19915-19924
Diffusion models trained on large-scale text-image datasets have demonstrated a strong capability of controllable high-quality image generation from arbitrary t ext prompts. However the generation quality and generalization ability of 3D dif

fusion models is hindered by the scarcity of high-quality and large-scale 3D dat asets. In this paper we present PI3D a framework that fully leverages the pre-tr ained text-to-image diffusion models' ability to generate high-quality 3D shapes from text prompts in minutes. The core idea is to connect the 2D and 3D domains by representing a 3D shape as a set of Pseudo RGB Images. We fine-tune an exist ing text-to-image diffusion model to produce such pseudo-images using a small nu mber of text-3D pairs. Surprisingly we find that it can already generate meaning ful and consistent 3D shapes given complex text descriptions. We further take th e generated shapes as the starting point for a lightweight iterative refinement using score distillation sampling to achieve high-quality generation under a low budget. PI3D generates a single 3D shape from text in only 3 minutes and the qu ality is validated to outperform existing 3D generative models by a large margin .
****
Orthogonal Adaptation for Modular Customization of Diffusion Models
Ryan Po, Guandao Yang, Kfir Aberman, Gordon Wetzstein; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7964-7973
Customization techniques for text-to-image models have paved the way for a wide range of previously unattainable applications enabling the generation of specifi c concepts across diverse contexts and styles. While existing methods facilitate high-fidelity customization for individual concepts or a limited pre-defined se t of them they fall short of achieving scalability where a single model can seam lessly render countless concepts. In this paper we address a new problem called Modular Customization with the goal of efficiently merging customized models tha t were fine-tuned independently for individual concepts. This allows the merged model to jointly synthesize concepts in one image without compromising fidelity or incurring any additional computational costs. To address this problem we intr oduce Orthogonal Adaptation a method designed to encourage the customized models which do not have access to each other during fine-tuning to have orthogonal re sidual weights. This ensures that during inference time the customized models ca n be summed with minimal interference. Our proposed method is both simple and ve rsatile applicable to nearly all optimizable weights in the model architecture. Through an extensive set of quantitative and qualitative evaluations our method consistently outperforms relevant baselines in terms of efficiency and identity preservation demonstrating a significant leap toward scalable customization of d iffusion models.
****
pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Re construction
David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, Vincent Sitzmann; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 19457-19467
We introduce pixelSplat a feed-forward model that learns to reconstruct 3D radia nce fields parameterized by 3D Gaussian primitives from pairs of images. Our mod el features real-time and memory-efficient rendering for scalable training as we ll as fast 3D reconstruction at inference time. To overcome local minima inheren t to sparse and locally supported representations we predict a dense probability distribution over 3D and sample Gaussian means from that probability distributi on. We make this sampling operation differentiable via a reparameterization tric k allowing us to back-propagate gradients through the Gaussian splatting represe ntation. We benchmark our method on wide-baseline novel view synthesis on the re al-world RealEstate10k and ACID datasets where we outperform state-of-the-art li ght field transformers and accelerate rendering by 2.5 orders of magnitude while reconstructing an interpretable and editable 3D radiance field. Additional mate rials can be found on the anonymous project website (pixelsplat.github.io).
****
VBench: Comprehensive Benchmark Suite for Video Generative Models
Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen

, Limin Wang, Dahua Lin, Yu Qiao, Ziwei Liu; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21807-21818
Video generation has witnessed significant advancements yet evaluating these mod
els remains a challenge. A comprehensive evaluation benchmark for video generati
on is indispensable for two reasons: 1) Existing metrics do not fully align with
 human perceptions; 2) An ideal evaluation system should provide insights to inf
orm future developments of video generation. To this end we present VBench a com
prehensive benchmark suite that dissects "video generation quality" into specifi
c hierarchical and disentangled dimensions each with tailored prompts and evalua
tion methods. VBench has three appealing properties: 1) Comprehensive Dimensions
: VBench comprises 16 dimensions in video generation (e.g. subject identity inco
nsistency motion smoothness temporal flickering and spatial relationship etc). T
he evaluation metrics with fine-grained levels reveal individual models' strengt
hs and weaknesses. 2) Human Alignment: We also provide a dataset of human prefer
ence annotations to validate our benchmarks' alignment with human perception for
 each evaluation dimension respectively. 3) Valuable Insights: We look into curr
ent models' ability across various evaluation dimensions and various content typ
es. We also investigate the gaps between video and image generation models. We w
ill open-source VBench including all prompts evaluation methods generated videos
 and human preference annotations and also include more video generation models
in VBench to drive forward the field of video generation.
****
Language-conditioned Detection Transformer
Jang Hyun Cho, Philipp Krähenbühl; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2024, pp. 16593-16603
We present a new open-vocabulary detection framework. Our framework uses both im
age-level labels and detailed detection annotations when available. Our framewor
k proceeds in three steps. We first train a language-conditioned object detector
 on fully-supervised detection data. This detector gets to see the presence or a
bsence of ground truth classes during training and conditions prediction on the
set of present classes. We use this detector to pseudo-label images with image-l
evel labels. Our detector provides much more accurate pseudo-labels than prior a
pproaches with its conditioning mechanism. Finally we train an unconditioned ope
n-vocabulary detector on the pseudo-annotated images. The resulting detector nam
ed DECOLA shows strong zero-shot performance in open-vocabulary LVIS benchmark a
s well as direct zero-shot transfer benchmarks on LVIS COCO Object365 and OpenIm
ages. DECOLA outperforms the prior arts by 17.1 AP-rare and 9.4 mAP on zero-shot
 LVIS benchmark. DECOLA achieves state-of-the-art results in various model sizes
 architectures and datasets by only training on open-sourced data and academic-s
cale computing. Code is available at https://github.com/janghyuncho/DECOLA.
****
Optimizing Diffusion Noise Can Serve As Universal Motion Priors
Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Su
wajanakorn, Siyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2024, pp. 1334-1345
We propose Diffusion Noise Optimization (DNO) a new method that effectively leve
rages existing motion diffusion models as motion priors for a wide range of moti
on-related tasks. Instead of training a task-specific diffusion model for each n
ew task DNO operates by optimizing the diffusion latent noise of an existing pre
-trained text-to-motion model. Given the corresponding latent noise of a human m
otion it propagates the gradient from the target criteria defined on the motion
space through the whole denoising process to update the diffusion latent noise.
As a result DNO supports any use cases where criteria can be defined as a functi
on of motion. In particular we show that for motion editing and control DNO outp
erforms existing methods in both achieving the objective and preserving the moti
on content. DNO accommodates a diverse range of editing modes including changing
 trajectory pose joint locations or avoiding newly added obstacles. In addition
DNO is effective in motion denoising and completion producing smooth and realist
ic motion from noisy and partial inputs. DNO achieves these results at inference
 time without the need for model retraining offering great versatility for any d

efined reward or loss function on the motion representation.
****
MAP: MAsk-Pruning for Source-Free Model Intellectual Property Protection
Boyang Peng, Sanqing Qu, Yong Wu, Tianpei Zou, Lianghua He, Alois Knoll, Guang Chen, Changjun Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23585-23594
Deep learning has achieved remarkable progress in various applications heightening the importance of safeguarding the intellectual property (IP) of well-trained models. It entails not only authorizing usage but also ensuring the deployment of models in authorized data domains i.e. making models exclusive to certain target domains. Previous methods necessitate concurrent access to source training data and target unauthorized data when performing IP protection making them risky and inefficient for decentralized private data. In this paper we target a practical setting where only a well-trained source model is available and investigate how we can realize IP protection. To achieve this we propose a novel MAsk Pruning (MAP) framework. MAP stems from an intuitive hypothesis i.e. there are target-related parameters in a well-trained model locating and pruning them is the key to IP protection. Technically MAP freezes the source model and learns a target-specific binary mask to prevent unauthorized data usage while minimizing performance degradation on authorized data. Moreover we introduce a new metric aimed at achieving a better balance between source and target performance degradation. To verify the effectiveness and versatility we have evaluated MAP in a variety of scenarios including vanilla source-available practical source-free and challenging data-free. Extensive experiments indicate that MAP yields new state-of-the-art performance.
****
Improving Single Domain-Generalized Object Detection: A Focus on Diversification and Alignment
Muhammad Sohail Danish, Muhammad Haris Khan, Muhammad Akhtar Munir, M. Saquib Sarfraz, Mohsen Ali; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17732-17742
In this work we tackle the problem of domain generalization for object detection specifically focusing on the scenario where only a single source domain is available. We propose an effective approach that involves two key steps: diversifying the source domain and aligning detections based on class prediction confidence and localization. Firstly we demonstrate that by carefully selecting a set of augmentations a base detector can outperform existing methods for single domain generalization by a good margin. This highlights the importance of domain diversification in improving the performance of object detectors. Secondly we introduce a method to align detections from multiple views considering both classification and localization outputs. This alignment procedure leads to better generalized and well-calibrated object detector models which are crucial for accurate decision-making in safety-critical applications. Our approach is detector-agnostic and can be seamlessly applied to both single-stage and two-stage detectors. To validate the effectiveness of our proposed methods we conduct extensive experiments and ablations on challenging domain-shift scenarios. The results consistently demonstrate the superiority of our approach compared to existing methods.
****
OVFoodSeg: Elevating Open-Vocabulary Food Image Segmentation via Image-Informed Textual Representation
Xiongwei Wu, Sicheng Yu, Ee-Peng Lim, Chong-Wah Ngo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4144-4153
In the realm of food computing segmenting ingredients from images poses substantial challenges due to the large intra-class variance among the same ingredients the emergence of new ingredients and the high annotation costs associated with large food segmentation datasets. Existing approaches primarily utilize a closed-vocabulary and static text embeddings setting. These methods often fall short in effectively handling the ingredients particularly new and diverse ones. In response to these limitations we introduce OVFoodSeg a framework that adopts an open

-vocabulary setting and enhances text embeddings with visual context. By integrating vision-language models (VLMs) our approach enriches text embedding with image-specific information through two innovative modules e.g. an image-to-text learner FoodLearner and an Image-Informed Text Encoder. The training process of OVFoodSeg is divided into two stages: the pre-training of FoodLearner and the subsequent learning phase for segmentation. The pre-training phase equips FoodLearner with the capability to align visual information with corresponding textual representations that are specifically related to food while the second phase adapts both the FoodLearner and the Image-Informed Text Encoder for the segmentation task. By addressing the deficiencies of previous models OVFoodSeg demonstrates a significant improvement achieving an 4.9% increase in mean Intersection over Union (mIoU) on the FoodSeg103 dataset setting a new milestone for food image segmentation.
****
XFeat: Accelerated Features for Lightweight Image Matching
Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, Erickson R. Nascimento; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2682-2691
We introduce a lightweight and accurate architecture for resource-efficient visual correspondence. Our method dubbed XFeat (Accelerated Features) revisits fundamental design choices in convolutional neural networks for detecting extracting and matching local features. Our new model satisfies a critical need for fast and robust algorithms suitable to resource-limited devices. In particular accurate image matching requires sufficiently large image resolutions -- for this reason we keep the resolution as large as possible while limiting the number of channels in the network. Besides our model is designed to offer the choice of matching at the sparse or semi-dense levels each of which may be more suitable for different downstream applications such as visual navigation and augmented reality. Our model is the first to offer semi-dense matching efficiently leveraging a novel match refinement module that relies on coarse local descriptors. XFeat is versatile and hardware-independent surpassing current deep learning-based local features in speed (up to 5x faster) with comparable or better accuracy proven in pose estimation and visual localization. We showcase it running in real-time on an inexpensive laptop CPU without specialized hardware optimizations. Code and weights are available at verlab.dcc.ufmg.br/descriptors/xfeat_cvpr24.
****
Visual Prompting for Generalized Few-shot Segmentation: A Multi-scale Approach
Mir Rayat Imtiaz Hossain, Mennatullah Siam, Leonid Sigal, James J. Little; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23470-23480
The emergence of attention-based transformer models has led to their extensive use in various tasks due to their superior generalization and transfer properties. Recent research has demonstrated that such models when prompted appropriately are excellent for few-shot inference. However such techniques are under-explored for dense prediction tasks like semantic segmentation. In this work we examine the effectiveness of prompting a transformer-decoder with learned visual prompts for the generalized few-shot segmentation (GFSS) task. Our goal is to achieve strong performance not only on novel categories with limited examples but also to retain performance on base categories. We propose an approach to learn visual prompts with limited examples. These learned visual prompts are used to prompt a multiscale transformer decoder to facilitate accurate dense predictions. Additionally we introduce a unidirectional causal attention mechanism between the novel prompts learned with limited examples and the base prompts learned with abundant data. This mechanism enriches the novel prompts without deteriorating the base class performance. Overall this form of prompting helps us achieve state-of-the-art performance for GFSS on two different benchmark datasets: COCO-20^i and Pascal-5^i without the need for test-time optimization (or transduction). Furthermore test-time optimization leveraging unlabelled test data can be used to improve the prompts which we refer to as transductive prompt tuning.
****

ARTrackV2: Prompting Autoregressive Tracker Where to Look and How to Describe
Yifan Bai, Zeyang Zhao, Yihong Gong, Xing Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19048-19057
We present ARTrackV2 which integrates two pivotal aspects of tracking: determining where to look (localization) and how to describe (appearance analysis) the target object across video frames. Building on the foundation of its predecessor ARTrackV2 extends the concept by introducing a unified generative framework to "read out" object's trajectory and "retell" its appearance in an autoregressive manner. This approach fosters a time-continuous methodology that models the joint evolution of motion and visual features guided by previous estimates. Furthermore ARTrackV2 stands out for its efficiency and simplicity obviating the less efficient intra-frame autoregression and hand-tuned parameters for appearance updates. Despite its simplicity ARTrackV2 achieves state-of-the-art performance on prevailing benchmark datasets while demonstrating a remarkable efficiency improvement. In particular ARTrackV2 achieves an AO score of 79. 5% on GOT-10k and an AUC of 86. 1% on TrackingNet while being 3.6 xfaster than ARTrack.
****
A Vision Check-up for Language Models
Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, Antonio Torralba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14410-14419
What does learning to model relationships between strings teach Large Language Models (LLMs) about the visual world? We systematically evaluate LLMs' abilities to generate and recognize an assortment of visual concepts of increasing complexity and then demonstrate how a preliminary visual representation learning system can be trained using models of text. As language models lack the ability to consume or output visual information as pixels we use code to represent images in our study. Although LLM-generated images do not look like natural images results on image generation and the ability of models to correct these generated images indicate that precise modeling of strings can teach language models about numerous aspects of the visual world. Furthermore experiments on self-supervised visual representation learning utilizing images generated with text models highlight the potential to train vision models capable of making semantic assessments of natural images using just LLMs.
****
Memory-based Adapters for Online 3D Scene Perception
Xiuwei Xu, Chong Xia, Ziwei Wang, Linqing Zhao, Yueqi Duan, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21604-21613
In this paper we propose a new framework for online 3D scene perception. Conventional 3D scene perception methods are offline i.e. take an already reconstructed 3D scene geometry as input which is not applicable in robotic applications where the input data is streaming RGB-D videos rather than a complete 3D scene reconstructed from pre- collected RGB-D videos. To deal with online 3D scene per- ception tasks where data collection and perception should be performed simultaneously the model should be able to process 3D scenes frame by frame and make use of the temporal information. To this end we propose an adapter-based plug-and-play module for the backbone of 3D scene perception model which constructs memory to cache and aggregate the extracted RGB-D features to empower offline models with temporal learning ability. Specifically we propose a queued memory mechanism to cache the supporting point cloud and image features. Then we devise aggregation modules which directly perform on the memory and pass temporal information to current frame. We further propose 3D-to-2D adapter to enhance image features with strong global context. Our adapters can be easily inserted into mainstream offline architectures of different tasks and significantly boost their performance on online tasks. Extensive experiments on ScanNet and SceneNN datasets demonstrate our approach achieves leading performance on three 3D scene perception tasks compared with state-of-the-art online methods by simply finetuning existing offline models without any model and task-specific designs.

****
SyncMask: Synchronized Attentional Masking for Fashion-centric Vision-Language Pretraining
Chull Hwan Song, Taebaek Hwang, Jooyoung Yoon, Shunghyun Choi, Yeong Hyeon Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13948-13957
Vision-language models (VLMs) have made significant strides in cross-modal understanding through large-scale paired datasets. However in fashion domain datasets often exhibit a disparity between the information conveyed in image and text. This issue stems from datasets containing multiple images of a single fashion item all paired with one text leading to cases where some textual details are not visible in individual images. This mismatch particularly when non-co-occurring elements are masked undermines the training of conventional VLM objectives like Masked Language Modeling and Masked Image Modeling thereby hindering the model's ability to accurately align fine-grained visual and textual features. Addressing this problem we propose Synchronized attentional Masking (SyncMask) which generate masks that pinpoint the image patches and word tokens where the information co-occur in both image and text. This synchronization is accomplished by harnessing cross-attentional features obtained from a momentum model ensuring a precise alignment between the two modalities. Additionally we enhance grouped batch sampling with semi-hard negatives effectively mitigating false negative issues in Image-Text Matching and Image-Text Contrastive learning objectives within fashion datasets. Our experiments demonstrate the effectiveness of the proposed approach outperforming existing methods in three downstream tasks.
****
A Study of Dropout-Induced Modality Bias on Robustness to Missing Video Frames for Audio-Visual Speech Recognition
Yusheng Dai, Hang Chen, Jun Du, Ruoyu Wang, Shihao Chen, Haotian Wang, Chin-Hui Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27445-27455
Advanced Audio-Visual Speech Recognition (AVSR) systems have been observed to be sensitive to missing video frames performing even worse than single-modality models. While applying the common dropout techniques to the video modality enhances robustness to missing frames it simultaneously results in a performance loss when dealing with complete data input. In this study we delve into this contrasting phenomenon through the lens of modality bias and uncover that an excessive modality bias towards the audio modality induced by dropout constitutes the fundamental cause. Next we present the Modality Bias Hypothesis (MBH) to systematically describe the relationship between the modality bias and the robustness against missing modality in multimodal systems. Building on these findings we propose a novel Multimodal Distribution Approximation with Knowledge Distillation (MDA-KD) framework to reduce over-reliance on the audio modality maintaining performance and robustness simultaneously. Finally to address an entirely missing modality we adopt adapters to dynamically switch decision strategies. The effectiveness of our proposed approach is evaluated through comprehensive experiments on the MISP2021 and MISP2022 datasets. Our code is available at https://github.com/dalision/ModalBiasAVSR.
****
A Conditional Denoising Diffusion Probabilistic Model for Point Cloud Upsampling
Wentao Qu, Yuantian Shao, Lingwu Meng, Xiaoshui Huang, Liang Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20786-20795
Point cloud upsampling (PCU) enriches the representation of raw point clouds significantly improving the performance in downstream tasks such as classification and reconstruction. Most of the existing point cloud upsampling methods focus on sparse point cloud feature extraction and upsampling module design. In a different way we dive deeper into directly modelling the gradient of data distribution from dense point clouds. In this paper we proposed a conditional denoising diffusion probabilistic model (DDPM) for point cloud upsampling called PUDM. Specifically PUDM treats the sparse point cloud as a condition and iteratively learns t

he transformation relationship between the dense point cloud and the noise. Simultaneously PUDM aligns with a dual mapping paradigm to further improve the discernment of point features. In this context PUDM enables learning complex geometry details in the ground truth through the dominant features while avoiding an additional upsampling module design. Furthermore to generate high-quality arbitrary-scale point clouds during inference PUDM exploits the prior knowledge of the scale between sparse point clouds and dense point clouds during training by parameterizing a rate factor. Moreover PUDM exhibits strong noise robustness in experimental results. In the quantitative and qualitative evaluations on PU1K and PUGAN PUDM significantly outperformed existing methods in terms of Chamfer Distance (CD) and Hausdorff Distance (HD) achieving state of the art (SOTA) performance.
****
VideoRF: Rendering Dynamic Radiance Fields as 2D Feature Video Streams
Liao Wang, Kaixin Yao, Chengcheng Guo, Zhirui Zhang, Qiang Hu, Jingyi Yu, Lan Xu, Minye Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 470-481
Neural Radiance Fields (NeRFs) excel in photorealistically rendering static scenes. However rendering dynamic long-duration radiance fields on ubiquitous devices remains challenging due to data storage and computational constraints. In this paper we introduce VideoRF the first approach to enable real-time streaming and rendering of dynamic human-centric radiance fields on mobile platforms. At the core is a serialized 2D feature image stream representing the 4D radiance field all in one. We introduce a tailored training scheme directly applied to this 2D domain to impose the temporal and spatial redundancy of the feature image stream. By leveraging the redundancy we show that the feature image stream can be efficiently compressed by 2D video codecs which allows us to exploit video hardware accelerators to achieve real-time decoding. On the other hand based on the feature image stream we propose a novel rendering pipeline for VideoRF which has specialized space mappings to query radiance properties efficiently. Paired with a deferred shading model VideoRF has the capability of real-time rendering on mobile devices thanks to its efficiency. We have developed a real-time interactive player that enables online streaming and rendering of dynamic scenes offering a seamless and immersive free-viewpoint experience across a range of devices from desktops to mobile phones. Our project page is available at https://aoliao12138.github.io/VideoRF/.
****
DPHMs: Diffusion Parametric Head Models for Depth-based Tracking
Jiapeng Tang, Angela Dai, Yinyu Nie, Lev Markhasin, Justus Thies, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1111-1122
We introduce Diffusion Parametric Head Models (DPHMs) a generative model that enables robust volumetric head reconstruction and tracking from monocular depth sequences. While recent volumetric head models such as NPHMs can now excel in representing high-fidelity head geometries tracking and reconstructing heads from real-world single-view depth sequences remains very challenging as the fitting to partial and noisy observations is underconstrained. To tackle these challenges we propose a latent diffusion-based prior to regularize volumetric head reconstruction and tracking. This prior-based regularizer effectively constrains the identity and expression codes to lie on the underlying latent manifold which represents plausible head shapes. To evaluate the effectiveness of the diffusion-based prior we collect a dataset of monocular Kinect sequences consisting of various complex facial expression motions and rapid transitions. We compare our method to state-of-the-art tracking methods and demonstrate improved head identity reconstruction as well as robust expression tracking.
****
DetDiffusion: Synergizing Generative and Perceptive Models for Enhanced Data Generation and Perception
Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, Kai Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 724

6-7255
Current perceptive models heavily depend on resource-intensive datasets promptin g the need for innovative solutions. Leveraging recent advances in diffusion mod els synthetic data by constructing image inputs from various annotations proves beneficial for downstream tasks. While prior methods have separately addressed g enerative and perceptive models DetDiffusion for the first time harmonizes both tackling the challenges in generating effective data for perceptive models. To e nhance image generation with perceptive models we introduce perception-aware los s (P.A. loss) through segmentation improving both quality and controllability. T o boost the performance of specific perceptive models our method customizes data augmentation by extracting and utilizing perception-aware attribute (P.A. Attr) during generation. Experimental results from the object detection task highligh t DetDiffusion's superior performance establishing a new state-of-the-art in lay out-guided generation. Furthermore image syntheses from DetDiffusion can effecti vely augment training data significantly enhancing downstream detection performa nce.
****
GAFusion: Adaptive Fusing LiDAR and Camera with Multiple Guidance for 3D Object Detection
Xiaotian Li, Baojie Fan, Jiandong Tian, Huijie Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21209-21 218
Recent years have witnessed the remarkable progress of 3D multi-modality object detection methods based on the Bird's-Eye-View (BEV) perspective. However most o f them overlook the complementary interaction and guidance between LiDAR and cam era. In this work we propose a novel multi-modality 3D objection detection metho d named GAFusion with LiDAR-guided global interaction and adaptive fusion. Speci fically we introduce sparse depth guidance (SDG) and LiDAR occupancy guidance (L OG) to generate 3D features with sufficient depth information. In the following LiDAR-guided adaptive fusion transformer (LGAFT) is developed to adaptively enha nce the interaction of different modal BEV features from a global perspective. M eanwhile additional downsampling with sparse height compression and multi-scale dual-path transformer (MSDPT) are designed to enlarge the receptive fields of di fferent modal features. Finally a temporal fusion module is introduced to aggreg ate features from previous frames. GAFusion achieves state-of-the-art 3D object detection results with 73.6% mAP and 74.9% NDS on the nuScenes test set.
****
Perception-Oriented Video Frame Interpolation via Asymmetric Blending
Guangyang Wu, Xin Tao, Changlin Li, Wenyi Wang, Xiaohong Liu, Qingqing Zheng; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2753-2762
Previous methods for Video Frame Interpolation (VFI) have encountered challenges notably the manifestation of blur and ghosting effects. These issues can be tra ced back to two pivotal factors: unavoidable motion errors and misalignment in s upervision. In practice motion estimates often prove to be error-prone resulting in misaligned features. Furthermore the reconstruction loss tends to bring blur ry results particularly in misaligned regions. To mitigate these challenges we p ropose a new paradigm called PerVFI (Perception-oriented Video Frame Interpolati on). Our approach incorporates an Asymmetric Synergistic Blending module (ASB) t hat utilizes features from both sides to synergistically blend intermediate feat ures. One reference frame emphasizes primary content while the other contributes complementary information. To impose a stringent constraint on the blending pro cess we introduce a self-learned sparse quasi-binary mask which effectively miti gates ghosting and blur artifacts in the output. Additionally we employ a normal izing flow-based generator and utilize the negative log-likelihood loss to learn the conditional distribution of the output which further facilitates the genera tion of clear and fine details. Experimental results validate the superiority of PerVFI demonstrating significant improvements in perceptual quality compared to existing methods. Codes are available at https://github.com/mulns/PerVFI
****

Countering Personalized Text-to-Image Generation with Influence Watermarks
Hanwen Liu, Zhicheng Sun, Yadong Mu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12257-12267
State-of-the-art personalized text-to-image generation systems are usually trained on a few reference images to learn novel visual representations. However this is likely to incur infringement of copyright for the reference image owners when these images are personal and publicly available. Recent progress has been made in protecting these images from unauthorized use by adding protective noises. Yet current protection methods work under the assumption that these protected images are not changed which is in contradiction to the fact that most public platforms intend to modify user-uploaded content e.g. image compression. This paper introduces a robust watermarking method namely InMark to protect images from unauthorized learning. Inspired by influence functions the proposed method forges protective watermarks on more important pixels for these reference images from both heuristic and statistical perspectives. In this way the personal semantics of these images are under protection even if these images are modified to some extent. Extensive experiments demonstrate that the proposed InMark outperforms previous state-of-the-art methods in both protective performance and robustness.
****
DUDF: Differentiable Unsigned Distance Fields with Hyperbolic Scaling
Miguel Fainstein, Viviana Siless, Emmanuel Iarussi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4484-4493
In recent years there has been a growing interest in training Neural Networks to approximate Unsigned Distance Fields (UDFs) for representing open surfaces in the context of 3D reconstruction. However UDFs are non-differentiable at the zero level set which leads to significant errors in distances and gradients generally resulting in fragmented and discontinuous surfaces. In this paper we propose to learn a hyperbolic scaling of the unsigned distance field which defines a new Eikonal problem with distinct boundary conditions. This allows our formulation to integrate seamlessly with state-of-the-art continuously differentiable implicit neural representation networks largely applied in the literature to represent signed distance fields. Our approach not only addresses the challenge of open surface representation but also demonstrates significant improvement in reconstruction quality and training performance. Moreover the unlocked field's differentiability allows the accurate computation of essential topological properties such as normal directions and curvatures pervasive in downstream tasks such as rendering. Through extensive experiments we validate our approach across various data sets and against competitive baselines. The results demonstrate enhanced accuracy and up to an order of magnitude increase in speed compared to previous methods.
****
PromptAD: Learning Prompts with only Normal Samples for Few-Shot Anomaly Detection
Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, Lizhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16838-16848
The vision-language model has brought great improvement to few-shot industrial anomaly detection which usually needs to design of hundreds of prompts through prompt engineering. For automated scenarios we first use conventional prompt learning with many-class paradigm as the baseline to automatically learn prompts but found that it can not work well in one-class anomaly detection. To address the above problem this paper proposes a one-class prompt learning method for few-shot anomaly detection termed PromptAD. First we propose semantic concatenation which can transpose normal prompts into anomaly prompts by concatenating normal prompts with anomaly suffixes thus constructing a large number of negative samples used to guide prompt learning in one-class setting. Furthermore to mitigate the training challenge caused by the absence of anomaly images we introduce the concept of explicit anomaly margin which is used to explicitly control the margin between normal prompt features and anomaly prompt features through a hyper-paramete

r. For image-level/pixel-level anomaly detection PromptAD achieves first place in 11/12 few-shot settings on MVTec and VisA.
****
Improving Graph Contrastive Learning via Adaptive Positive Sampling
Jiaming Zhuo, Feiyang Qin, Can Cui, Kun Fu, Bingxin Niu, Mengzhu Wang, Yuanfang Guo, Chuan Wang, Zhen Wang, Xiaochun Cao, Liang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23179-23187
Graph Contrastive Learning (GCL) a Self-Supervised Learning (SSL) architecture tailored for graphs has shown notable potential for mitigating label scarcity. Its core idea is to amplify feature similarities between the positive sample pairs and reduce them between the negative sample pairs. Unfortunately most existing GCLs consistently present suboptimal performances on both homophilic and heterophilic graphs. This is primarily attributed to two limitations of positive sampling that is incomplete local sampling and blind sampling. To address these limitations this paper introduces a novel GCL framework with an adaptive positive sampling module named grapH contrastivE Adaptive posiTive Samples (HEATS). Motivated by the observation that the affinity matrix corresponding to optimal positive sample sets has a block-diagonal structure with equal weights within each block a self-expressive learning objective incorporating the block and idempotent constraint is presented. This learning objective and the contrastive learning objective are iteratively optimized to improve the adaptability and robustness of HEATS. Extensive experiments on graphs and images validate the effectiveness and generality of HEATS.
****
UFC-Net: Unrolling Fixed-point Continuous Network for Deep Compressive Sensing
Xiaoyang Wang, Hongping Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25149-25159
Deep unfolding networks (DUNs) renowned for their interpretability and superior performance have invigorated the realm of compressive sensing (CS). Nonetheless existing DUNs frequently suffer from issues related to insufficient feature extraction and feature attrition during the iterative steps. In this paper we propose Unrolling Fixed-point Continuous Network (UFC-Net) a novel deep CS framework motivated by the traditional fixed-point continuous optimization algorithm. Specifically we introduce Convolution-guided Attention Module (CAM) to serve as a critical constituent within the reconstruction phase encompassing tailored components such as Multi-head Attention Residual Block (MARB) Auxiliary Iterative Reconstruction Block (AIRB) etc. MARB effectively integrates multi-head attention mechanisms with convolution to reinforce feature extraction transcending the confinement of localized attributes and facilitating the apprehension of long-range correlations. Meanwhile AIRB introduces auxiliary variables significantly bolstering the preservation of features within each iterative stage. Extensive experiments demonstrate that our proposed UFC-Net achieves remarkable performance both on image CS and CS-magnetic resonance imaging (CS-MRI) in contrast to state-of-the-art methods.
****
ECoDepth: Effective Conditioning of Diffusion Models for Monocular Depth Estimation
Suraj Patni, Aradhye Agarwal, Chetan Arora; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28285-28295
In the absence of parallax cues a learning-based single image depth estimation (SIDE) model relies heavily on shading and contextual cues in the image. While this simplicity is attractive it is necessary to train such models on large and varied datasets which are difficult to capture. It has been shown that using embeddings from pre-trained foundational models such as CLIP improves zero shot transfer in several applications. Taking inspiration from this in our paper we explore the use of global image priors generated from a pre-trained ViT model to provide more detailed contextual information. We argue that the embedding vector from a ViT model pre-trained on a large dataset captures greater relevant information for SIDE than the usual route of generating pseudo image captions followed by

CLIP based text embeddings. Based on this idea we propose a new SIDE model using a diffusion backbone which is conditioned on ViT embeddings. Our proposed design establishes a new state-of-the-art (SOTA) for SIDE on NYUv2 dataset achieving Abs Rel error of 0.059(14% improvement) compared to 0.069 by the current SOTA (VPD). And on KITTI dataset achieving Sq Rel error of 0.139 (2% improvement) compared to 0.142 by the current SOTA (GEDepth). For zero-shot transfer with a model trained on NYUv2 we report mean relative improvement of (20% 23% 81% 25%) over NeWCRFs on (Sun-RGBD iBims1 DIODE HyperSim) datasets compared to (16% 18% 45% 9%) by ZoeDepth. The project page is available at https://ecodepth-iitd.github.io
****
DL3DV-10K: A Large-Scale Scene Dataset for Deep Learning-based 3D Vision
Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, Aniket Bera; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22160-22169
We have witnessed significant progress in deep learning-based 3D vision ranging from neural radiance field (NeRF) based 3D representation learning to applications in novel view synthesis (NVS). However existing scene-level datasets for deep learning-based 3D vision limited to either synthetic environments or a narrow selection of real-world scenes are quite insufficient. This insufficiency not only hinders a comprehensive benchmark of existing methods but also caps what could be explored in deep learning-based 3D analysis. To address this critical gap we present DL3DV-10K a large-scale scene dataset featuring 51.2 million frames from 10510 videos captured from 65 types of point-of-interest (POI) locations covering both bounded and unbounded scenes with different levels of reflection transparency and lighting. We conducted a comprehensive benchmark of recent NVS methods on DL3DV-10K which revealed valuable insights for future research in NVS. In addition we have obtained encouraging results in a pilot study to learn generalizable NeRF from DL3DV-10K which manifests the necessity of a large-scale scene-level dataset to forge a path toward a foundation model for learning 3D representation. Our DL3DV-10K dataset benchmark results and models will be publicly accessible.
****
2S-UDF: A Novel Two-stage UDF Learning Method for Robust Non-watertight Model Reconstruction from Multi-view Images
Junkai Deng, Fei Hou, Xuhui Chen, Wencheng Wang, Ying He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5084-5093
Recently building on the foundation of neural radiance field various techniques have emerged to learn unsigned distance fields (UDF) to reconstruct 3D non-watertight models from multi-view images. Yet a central challenge in UDF-based volume rendering is formulating a proper way to convert unsigned distance values into volume density ensuring that the resulting weight function remains unbiased and sensitive to occlusions. Falling short on these requirements often results in incorrect topology or large reconstruction errors in resulting models. This paper addresses this challenge by presenting a novel two-stage algorithm 2S-UDF for learning a high-quality UDF from multi-view images. Initially the method applies an easily trainable density function that while slightly biased and transparent aids in coarse reconstruction. The subsequent stage then refines the geometry and appearance of the object to achieve a high-quality reconstruction by directly adjusting the weight function used in volume rendering to ensure that it is unbiased and occlusion-aware. Decoupling density and weight in two stages makes our training stable and robust distinguishing our technique from existing UDF learning approaches. Evaluations on the DeepFashion3D DTU and BlendedMVS datasets validate the robustness and effectiveness of our proposed approach. In both quantitative metrics and visual quality the results indicate our superior performance over other UDF learning techniques in reconstructing 3D non-watertight models from multi-view images. Our code is available at https://bitbucket.org/jkdeng/2sudf/.
****

DETRs Beat YOLOs on Real-time Object Detection
Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, Jie Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16965-16974
The YOLO series has become the most popular framework for real-time object detection due to its reasonable trade-off between speed and accuracy. However we observe that the speed and accuracy of YOLOs are negatively affected by the NMS. Recently end-to-end Transformer-based detectors (DETRs) have provided an alternative to eliminating NMS. Nevertheless the high computational cost limits their practicality and hinders them from fully exploiting the advantage of excluding NMS. In this paper we propose the Real-Time DEtection TRansformer (RT-DETR) the first real-time end-to-end object detector to our best knowledge that addresses the above dilemma. We build RT-DETR in two steps drawing on the advanced DETR: first we focus on maintaining accuracy while improving speed followed by maintaining speed while improving accuracy. Specifically we design an efficient hybrid encoder to expeditiously process multi-scale features by decoupling intra-scale interaction and cross-scale fusion to improve speed. Then we propose the uncertainty-minimal query selection to provide high-quality initial queries to the decoder thereby improving accuracy. In addition RT-DETR supports flexible speed tuning by adjusting the number of decoder layers to adapt to various scenarios without retraining. Our RT-DETR-R50 / R101 achieves 53.1% / 54.3% AP on COCO and 108 / 74 FPS on T4 GPU outperforming previously advanced YOLOs in both speed and accuracy. We also develop scaled RT-DETRs that outperform the lighter YOLO detectors (S and M models). Furthermore RT-DETR-R50 outperforms DINO-R50 by 2.2% AP in accuracy and about 21 times in FPS. After pre-training with Objects365 RT-DETR-R50 / R101 achieves 55.3% / 56.2% AP. The project page: https://zhao-yian.github.io/RTDETR.
****
UniVS: Unified and Universal Video Segmentation with Prompts as Queries
Minghan Li, Shuai Li, Xindong Zhang, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3227-3238
Despite the recent advances in unified image segmentation (IS) developing a unified video segmentation (VS) model remains a challenge. This is mainly because generic category-specified VS tasks need to detect all objects and track them across consecutive frames while prompt-guided VS tasks require re-identifying the target with visual/text prompts throughout the entire video making it hard to handle the different tasks with the same architecture. We make an attempt to address these issues and present a novel unified VS architecture namely UniVS by using prompts as queries. UniVS averages the prompt features of the target from previous frames as its initial query to explicitly decode masks and introduces a target-wise prompt cross-attention layer in the mask decoder to integrate prompt features in the memory pool. By taking the predicted masks of entities from previous frames as their visual prompts UniVS converts different VS tasks into prompt-guided target segmentation eliminating the heuristic inter-frame matching process. Our framework not only unifies the different VS tasks but also naturally achieves universal training and testing ensuring robust performance across different scenarios. UniVS shows a commendable balance between performance and universality on 10 challenging VS benchmarks covering video instance semantic panoptic object and referring segmentation tasks. Code can be found at https://github.com/MinghanLi/UniVS.
****
Bilateral Adaptation for Human-Object Interaction Detection with Occlusion-Robustness
Guangzhi Wang, Yangyang Guo, Ziwei Xu, Mohan Kankanhalli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27970-27980
Human-Object Interaction (HOI) Detection constitutes an important aspect of human-centric scene understanding which requires precise object detection and interaction recognition. Despite increasing advancement in detection recognizing subtle and intricate interactions remains challenging. Recent methods have endeavored

 to leverage the rich semantic representation from pre-trained CLIP yet fail to efficiently capture finer-grained spatial features that are highly informative for interaction discrimination. In this work instead of solely using representations from CLIP we fill the gap by proposing a spatial adapter that efficiently utilizes the multi-scale spatial information in the pre-trained detector. This leads to a bilateral adaptation that mutually produces complementary features. To further improve interaction recognition under occlusion which is common in crowded scenarios we propose an Occluded Part Extrapolation module that guides the model to recover the spatial details from manually occluded feature maps. Moreover we design a Conditional Contextual Mining module that further mines informative contextual clues from the spatial features via a tailored cross-attention mechanism. Extensive experiments on V-COCO and HICO-DET benchmarks demonstrate that our method significantly outperforms prior art on both standard and zero-shot settings resulting in new state-of-the-art performance. Additional ablation studies further validate the effectiveness of each component in our method.
****
An Asymmetric Augmented Self-Supervised Learning Method for Unsupervised Fine-Grained Image Hashing
Feiran Hu, Chenlin Zhang, Jiangliang Guo, Xiu-Shen Wei, Lin Zhao, Anqi Xu, Lingyan Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17648-17657
Unsupervised fine-grained image hashing aims to learn compact binary hash codes in unsupervised settings addressing challenges posed by large-scale datasets and dependence on supervision. In this paper we first identify a granularity gap between generic and fine-grained datasets for unsupervised hashing methods highlighting the inadequacy of conventional self-supervised learning for fine-grained visual objects. To bridge this gap we propose the Asymmetric Augmented Self-Supervised Learning (A^2-SSL) method comprising three modules. The asymmetric augmented SSL module employs suitable augmentation strategies for positive/negative views preventing fine-grained category confusion inherent in conventional SSL. Part-oriented dense contrastive learning utilizes the Fisher Vector framework to capture and model fine-grained object parts enhancing unsupervised representations through part-level dense contrastive learning. Self-consistent hash code learning introduces a reconstruction task aligned with the self-consistency principle guiding the model to emphasize comprehensive features particularly fine-grained patterns. Experimental results on five benchmark datasets demonstrate the superiority of A^2-SSL over existing methods affirming its efficacy in unsupervised fine-grained image hashing.
****
Efficiently Assemble Normalization Layers and Regularization for Federated Domain Generalization
Khiem Le, Long Ho, Cuong Do, Danh Le-Phuoc, Kok-Seng Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6027-6036
Domain shift is a formidable issue in Machine Learning that causes a model to suffer from performance degradation when tested on unseen domains. Federated Domain Generalization (FedDG) attempts to train a global model using collaborative clients in a privacy-preserving manner that can generalize well to unseen clients possibly with domain shift. However most existing FedDG methods either cause additional privacy risks of data leakage or induce significant costs in client communication and computation which are major concerns in the Federated Learning paradigm. To circumvent these challenges here we introduce a novel architectural method for FedDG namely gPerXAN which relies on a normalization scheme working with a guiding regularizer. In particular we carefully design Personalized eXplicitly Assembled Normalization to enforce client models selectively filtering domain-specific features that are biased towards local data while retaining discrimination of those features. Then we incorporate a simple yet effective regularizer to guide these models in directly capturing domain-invariant representations that the global model's classifier can leverage. Extensive experimental results on two benchmark datasets i.e. PACS and Office-Home and a real-world medical dataset

Camelyon17 indicate that our proposed method outperforms other existing methods in addressing this particular problem.
****
Exploring Pose-Aware Human-Object Interaction via Hybrid Learning
Eastman Z Y Wu, Yali Li, Yuan Wang, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17815-17825
Human-Object Interaction (HOI) detection plays a crucial role in visual scene comprehension. In recent advancements two-stage detectors have taken a prominent position. However they are encumbered by two primary challenges. First the misalignment between feature representation and relation reasoning gives rise to a deficiency in discriminative features crucial for interaction detection. Second due to sparse annotation the second-stage interaction head generates numerous candidate <human object> pairs with only a small fraction receiving supervision. Towards these issues we propose a hybrid learning method based on pose-aware HOI feature refinement. Specifically we devise pose-aware feature refinement that encodes spatial features by considering human body pose characteristics. It can direct attention towards key regions ultimately offering a wealth of fine-grained features imperative for HOI detection. Further we introduce a hybrid learning method that combines HOI triplets with probabilistic soft labels supervision which is regenerated from decoupled verb-object pairs. This method explores the implicit connections between the interactions enhancing model generalization without requiring additional data. Our method establishes state-of-the-art performance on HICO-DET benchmark and excels notably in detecting rare HOIs.
****
Depth Information Assisted Collaborative Mutual Promotion Network for Single Image Dehazing
Yafei Zhang, Shen Zhou, Huafeng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2846-2855
Recovering a clear image from a single hazy image is an open inverse problem. Although significant research progress has been made most existing methods ignore the effect that downstream tasks play in promoting upstream dehazing. From the perspective of the haze generation mechanism there is a potential relationship between the depth information of the scene and the hazy image. Based on this we propose a dual-task collaborative mutual promotion framework to achieve the dehazing of a single image. This framework integrates depth estimation and dehazing by a dual-task interaction mechanism and achieves mutual enhancement of their performance. To realize the joint optimization of the two tasks an alternative implementation mechanism with the difference perception is developed. On the one hand the difference perception between the depth maps of the dehazing result and the ideal image is proposed to promote the dehazing network to pay attention to the non-ideal areas of the dehazing. On the other hand by improving the depth estimation performance in the difficult-to-recover areas of the hazy image the dehazing network can explicitly use the depth information of the hazy image to assist the clear image recovery. To promote the depth estimation we propose to use the difference between the dehazed image and the ground truth to guide the depth estimation network to focus on the dehazed unideal areas. It allows dehazing and depth estimation to leverage their strengths in a mutually reinforcing manner. Experimental results show that the proposed method can achieve better performance than that of the state-of-the-art approaches. The source code is released at https://github.com/zhoushen1/DCMPNet.
****
Density-Adaptive Model Based on Motif Matrix for Multi-Agent Trajectory Prediction
Di Wen, Haoran Xu, Zhaocheng He, Zhe Wu, Guang Tan, Peixi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14822-14832
Multi-agent trajectory prediction is essential in autonomous driving risk avoidance and traffic flow control. However the heterogeneous traffic density on interactions which caused by physical laws social norms and so on is often overlooked

in existing methods. When the density varies the number of agents involved in interactions and the corresponding interaction probability change dynamically. To tackle this issue we propose a new method called \underline D ensity-\underline A daptive Model based on \underline M otif \underline M atrix for Multi-Agent Trajectory Prediction (DAMM) to gain insights into multi-agent systems. Here we leverage the motif matrix to represent dynamic connectivity in a higher-order pattern and distill the interaction information from the perspectives of the spatial and the temporal dimensions. Specifically in spatial dimension we utilize multi-scale feature fusion to adaptively select the optimal range of neighbors participating in interactions for each time slot. In temporal dimension we extract the temporal interaction features and adapt a pyramidal pooling layer to generate the interaction probability for each agent. Experimental results demonstrate that our approach surpasses state-of-the-art methods on autonomous driving dataset.
****
Contrastive Learning for DeepFake Classification and Localization via Multi-Label Ranking
Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17627-17637
We propose a unified approach to simultaneously addressing the conventional setting of binary deepfake classification and a more challenging scenario of uncovering what facial components have been forged as well as the exact order of the manipulations. To solve the former task we consider multiple instance learning (MIL) that takes each image as a bag and its patches as instances. A positive bag corresponds to a forged image that includes at least one manipulated patch (i.e. a pixel in the feature map). The formulation allows us to estimate the probability of an input image being a fake one and establish the corresponding contrastive MIL loss. On the other hand tackling the component-wise deepfake problem can be reduced to solving multi-label prediction but the requirement to recover the manipulation order further complicates the learning task into a multi-label ranking problem. We resolve this difficulty by designing a tailor-made loss term to enforce that the rank order of the predicted multi-label probabilities respects the ground-truth order of the sequential modifications of a deepfake image. Through extensive experiments and comparisons with other relevant techniques we provide extensive results and ablation studies to demonstrate that the proposed method is an overall more comprehensive solution to deepfake detection.
****
Unlocking the Potential of Pre-trained Vision Transformers for Few-Shot Semantic Segmentation through Relationship Descriptors
Ziqin Zhou, Hai-Ming Xu, Yangyang Shu, Lingqiao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3817-3827
The recent advent of pre-trained vision transformers has unveiled a promising property: their inherent capability to group semantically related visual concepts. In this paper we explore to harnesses this emergent feature to tackle few-shot semantic segmentation a task focused on classifying pixels in a test image with a few example data. A critical hurdle in this endeavor is preventing overfitting to the limited classes seen during training the few-shot segmentation model. As our main discovery we find that the concept of "relationship descriptors" initially conceived for enhancing the CLIP model for zero-shot semantic segmentation offers a potential solution. We adapt and refine this concept to craft a relationship descriptor construction tailored for few-shot semantic segmentation extending its application across multiple layers to enhance performance. Building upon this adaptation we proposed a few-shot semantic segmentation framework that is not only easy to implement and train but also effectively scales with the number of support examples and categories. Through rigorous experimentation across various datasets including PASCAL-5^ i and COCO-20^ i we demonstrate a clear advantage of our method in diverse few-shot semantic segmentation scenarios and a range of pre-trained vision transformer models. The findings clearly show that our method significantly outperforms current state-of-the-art techniques highlighting the effectiveness of harnessing the emerging capabilities of vision transform

ers for few-shot semantic segmentation. We release the code at https://github.com/ZiqinZhou66/FewSegwithRD.git.
****
CustomListener: Text-guided Responsive Interaction for User-friendly Listening Head Generation
Xi Liu, Ying Guo, Cheng Zhen, Tong Li, Yingying Ao, Pengfei Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2415-2424
Listening head generation aims to synthesize a non-verbal responsive listener head by modeling the correlation between the speaker and the listener in dynamic conversion. The applications of listener agent generation in virtual interaction have promoted many works achieving diverse and fine-grained motion generation. However they can only manipulate motions through simple emotional labels but cannot freely control the listener's motions. Since listener agents should have human-like attributes (e.g. identity personality) which can be freely customized by users this limits their realism. In this paper we propose a user-friendly framework called CustomListener to realize the free-form text prior guided listener generation. To achieve speaker-listener coordination we design a Static to Dynamic Portrait module (SDP) which interacts with speaker information to transform static text into dynamic portrait token with completion rhythm and amplitude information. To achieve coherence between segments we design a Past Guided Generation module (PGG) to maintain the consistency of customized listener attributes through the motion prior and utilize a diffusion-based structure conditioned on the portrait token and the motion prior to realize the controllable generation. To train and evaluate our model we have constructed two text-annotated listening head datasets based on ViCo and RealTalk which provide text-video paired labels. Extensive experiments have verified the effectiveness of our model.
****
Projecting Trackable Thermal Patterns for Dynamic Computer Vision
Mark Sheinin, Aswin C. Sankaranarayanan, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25223-25232
Adding artificial patterns to objects like QR codes can ease tasks such as object tracking robot navigation and conveying information (e.g. a label or a website link). However these patterns require a physical application and they alter the object's appearance. Conversely projected patterns can temporarily change the object's appearance aiding tasks like 3D scanning and retrieving object textures and shading. However projected patterns impede dynamic tasks like object tracking because they do not `stick' to the object's surface. Or do they? This paper introduces a novel approach combining the advantages of projected and persistent physical patterns. Our system projects heat patterns using a laser beam (similar in spirit to a LIDAR) which a thermal camera observes and tracks. Such thermal patterns enable tracking poorly-textured objects whose tracking is highly challenging with standard cameras while not affecting the object's appearance or physical properties. To avail these thermal patterns in existing vision frameworks we train a network to reverse heat diffusion's effects and remove inconsistent pattern points between different thermal frames. We prototyped and tested this approach on dynamic vision tasks like structure from motion optical flow and object tracking of everyday textureless objects.
****
SG-PGM: Partial Graph Matching Network with Semantic Geometric Fusion for 3D Scene Graph Alignment and Its Downstream Tasks
Yaxu Xie, Alain Pagani, Didier Stricker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28401-28411
Scene graphs have been recently introduced into 3D spatial understanding as a comprehensive representation of the scene. The alignment between 3D scene graphs is the first step of many downstream tasks such as scene graph aided point cloud registration mosaicking overlap checking and robot navigation. In this work we treat 3D scene graph alignment as a partial graph-matching problem and propose to solve it with a graph neural network. We reuse the geometric features learned b

y a point cloud registration method and associate the clustered point-level geometric features with the node-level semantic feature via our designed feature fusion module. Partial matching is enabled by using a learnable method to select the top-k similar node pairs. Subsequent downstream tasks such as point cloud registration are achieved by running a pre-trained registration network within the matched regions. We further propose a point-matching rescoring method that uses the node-wise alignment of the 3D scene graph to reweight the matching candidates from a pre-trained point cloud registration method. It reduces the false point correspondences estimated especially in low-overlapping cases. Experiments show that our method improves the alignment accuracy by 10 20% in low-overlap and random transformation scenarios and outperforms the existing work in multiple downstream tasks. Our code and models are available here (https://github.com/dfki-av/sg-pgm.git).
****
Fun with Flags: Robust Principal Directions via Flag Manifolds
Nathan Mankovich, Gustau Camps-Valls, Tolga Birdal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 330-340
Principal component analysis (PCA) along with its extensions to manifolds and outlier contaminated data have been indispensable in computer vision and machine learning. In this work we present a unifying formalism for PCA and its variants and introduce a framework based on the flags of linear subspaces i.e. a hierarchy of nested linear subspaces of increasing dimension which not only allows for a common implementation but also yields novel variants not explored previously. We begin by generalizing traditional PCA methods that either maximize variance or minimize reconstruction error. We expand these interpretations to develop a wide array of new dimensionality reduction algorithms by accounting for outliers and the data manifold. To devise a common computational approach we recast robust and dual forms of PCA as optimization problems on flag manifolds. We then integrate tangent space approximations of principal geodesic analysis (tangent-PCA) into this flag-based framework creating novel robust and dual geodesic PCA variations. The remarkable flexibility offered by the `flagification' introduced here enables even more algorithmic variants identified by specific flag types. Last but not least we propose an effective convergent solver for these flag-formulations employing the Stiefel manifold. Our empirical results on both real-world and synthetic scenarios demonstrate the superiority of our novel algorithms especially in terms of robustness to outliers on manifolds.
****
Generating Non-Stationary Textures using Self-Rectification
Yang Zhou, Rongjun Xiao, Dani Lischinski, Daniel Cohen-Or, Hui Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7767-7776
This paper addresses the challenge of example-based non-stationary texture synthesis. We introduce a novel two-step approach wherein users first modify a reference texture using standard image editing tools yielding an initial rough target for the synthesis. Subsequently our proposed method termed "self-rectification" automatically refines this target into a coherent seamless texture while faithfully preserving the distinct visual characteristics of the reference exemplar. Our method leverages a pre-trained diffusion network and uses self-attention mechanisms to gradually align the synthesized texture with the reference ensuring the retention of the structures in the provided target. Through experimental validation our approach exhibits exceptional proficiency in handling non-stationary textures demonstrating significant advancements in texture synthesis when compared to existing state-of-the-art techniques. Code is available at https://github.com/xiaorongjun000/Self-Rectification
****
SPU-PMD: Self-Supervised Point Cloud Upsampling via Progressive Mesh Deformation
Yanzhe Liu, Rong Chen, Yushi Li, Yixi Li, Xuehou Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5188-5197
Despite the success of recent upsampling approaches generating high-resolution p

oint sets with uniform distribution and meticulous structures is still challenging. Unlike existing methods that only take spatial information of the raw data into account we regard point cloud upsampling as generating dense point clouds from deformable topology. Motivated by this we present SPU-PMD a self-supervised topological mesh deformation network for 3D densification. As a cascaded framework our architecture is formulated by a series of coarse mesh interpolator and mesh deformers. At each stage the mesh interpolator first produces the initial dense point clouds via mesh interpolation which allows the model to perceive the primitive topology better. Meanwhile the deformer infers the morphing by estimating the movements of mesh nodes and reconstructs the descriptive topology structure. By associating mesh deformation with feature expansion this module progressively refines point clouds' surface uniformity and structural details. To demonstrate the effectiveness of the proposed method extensive quantitative and qualitative experiments are conducted on synthetic and real-scanned 3D data. Also we compare it with state-of-the-art techniques to further illustrate the superiority of our network. The project page is: https://github.com/lyz21/SPU-PMD
****
Advancing Saliency Ranking with Human Fixations: Dataset Models and Benchmarks
Bowen Deng, Siyang Song, Andrew P. French, Denis Schluppeck, Michael P. Pound; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28348-28357
Saliency ranking detection (SRD) has emerged as a challenging task in computer vision aiming not only to identify salient objects within images but also to rank them based on their degree of saliency. Existing SRD datasets have been created primarily using mouse-trajectory data which inadequately captures the intricacies of human visual perception. Addressing this gap this paper introduces the first large-scale SRD dataset SIFR constructed using genuine human fixation data thereby aligning more closely with real visual perceptual processes. To establish a baseline for this dataset we propose QAGNet a novel model that leverages salient instance query features from a transformer detector within a tri-tiered nested graph. Through extensive experiments we demonstrate that our approach outperforms existing state-of-the-art methods across two widely used SRD datasets and our newly proposed dataset. Code and dataset are available at https://github.com/EricDengbowen/QAGNet.
****
Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis
Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, Sergey Tulyakov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7038-7048
Contemporary models for generating images show remarkable quality and versatility. Swayed by these advantages the research community repurposes them to generate videos. Since video content is highly redundant we argue that naively bringing advances of image models to the video generation domain reduces motion fidelity visual quality and impairs scalability. In this work we build Snap Video a video-first model that systematically addresses these challenges. To do that we first extend the EDM framework to take into account spatially and temporally redundant pixels and naturally support video generation. Second we show that a U-Net--a workhorse behind image generation--scales poorly when generating videos requiring significant computational overhead. Hence we propose a new transformer-based architecture that trains 3.31 times faster than U-Nets (and is 4.5 faster at inference). This allows us to efficiently train a text-to-video model with billions of parameters for the first time reach state-of-the-art results on a number of benchmarks and generate videos with substantially higher quality temporal consistency and motion complexity. The user studies showed that our model was favored by a large margin over the most recent methods.
****
Unsupervised Deep Unrolling Networks for Phase Unwrapping
Zhile Chen, Yuhui Quan, Hui Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25182-25192

Phase unwrapping (PU) is a technique to reconstruct original phase images from their noisy wrapped counterparts finding many applications in scientific imaging. Although supervised learning has shown promise in PU its utility is limited in ground-truth (GT) scarce scenarios. This paper presents an unsupervised learning approach that eliminates the need for GTs during end-to-end training. Our approach leverages the insight that both the gradients and wrapped gradients of wrapped phases serve as noisy labels for GT phase gradients along with sparse outliers induced by the wrapping operation. A recorruption-based self-reconstruction loss in the gradient domain is proposed to mitigate the adverse effects of label noise complemented with a self-distillation loss for improved generalization. Additionally by unfolding a variational model of PU that utilizes wrapped gradients of wrapped phases for its data-fitting term we develop a deep unrolling network that encodes physics of phase wrapping and incorporates special treatments on outliers. In the experiments on three types of phase data our approach outperforms existing GT-free methods and competes well against the supervised ones.
****
Federated Generalized Category Discovery
Nan Pu, Wenjing Li, Xingyuan Ji, Yalan Qin, Nicu Sebe, Zhun Zhong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28741-28750
Generalized category discovery (GCD) aims at grouping unlabeled samples from known and unknown classes given labeled data of known classes. To meet the recent decentralization trend in the community we introduce a practical yet challenging task Federated GCD (Fed-GCD) where the training data are distributed in local clients and cannot be shared among clients. Fed-GCD aims to train a generic GCD model by client collaboration under the privacy-protected constraint. The Fed-GCD leads to two challenges: 1) representation degradation caused by training each client model with fewer data than centralized GCD learning and 2) highly heterogeneous label spaces across different clients. To this end we propose a novel Associated Gaussian Contrastive Learning (AGCL) framework based on learnable GMMs which consists of a Client Semantics Association (CSA) and a global-local GMM Contrastive Learning (GCL). On the server CSA aggregates the heterogeneous categories of local-client GMMs to generate a global GMM containing more comprehensive category knowledge. On each client GCL builds class-level contrastive learning with both local and global GMMs. The local GCL learns robust representation with limited local data. The global GCL encourages the model to produce more discriminative representation with the comprehensive category relationships that may not exist in local data. We build a benchmark based on six visual datasets to facilitate the study of Fed-GCD. Extensive experiments show that our AGCL outperforms multiple baselines on all datasets.
****
JointSQ: Joint Sparsification-Quantization for Distributed Learning
Weiying Xie, Haowei Li, Jitao Ma, Yunsong Li, Jie Lei, Donglai Liu, Leyuan Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5778-5787
Gradient sparsification and quantization offer a promising prospect to alleviate the communication overhead problem in distributed learning. However direct combination of the two results in suboptimal solutions due to the fact that sparsification and quantization haven't been learned together. In this paper we propose Joint Sparsification-Quantization (JointSQ) inspired by the discovery that sparsification can be treated as 0-bit quantization regardless of architectures. Specifically we mathematically formulate JointSQ as a mixed-precision quantization problem expanding the solution space. It can be solved by the designed MCKP-Greedy algorithm. Theoretical analysis demonstrates the minimal compression noise of JointSQ and extensive experiments on various network architectures including CNN RNN and Transformer also validate this point. Under the introduction of computation overhead consistent with or even lower than previous methods JointSQ achieves a compression ratio of 1000xon different models while maintaining near-lossless accuracy and brings 1.4xto 2.9xspeedup over existing methods.
****

A Unified Framework for Human-centric Point Cloud Video Understanding
Yiteng Xu, Kecheng Ye, Xiao Han, Yiming Ren, Xinge Zhu, Yuexin Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1155-1164
Human-centric Point Cloud Video Understanding (PVU) is an emerging field focused on extracting and interpreting human-related features from sequences of human point clouds further advancing downstream human-centric tasks and applications. Previous works usually focus on tackling one specific task and rely on huge labeled data which has poor generalization capability. Considering that human has specific characteristics including the structural semantics of human body and the dynamics of human motions we propose a unified framework to make full use of the prior knowledge and explore the inherent features in the data itself for generalized human-centric point cloud video understanding. Extensive experiments demonstrate that our method achieves state-of-the-art performance on various human-related tasks including action recognition and 3D pose estimation. All datasets and code will be released soon.
****
Edge-Aware 3D Instance Segmentation Network with Intelligent Semantic Prior
Wonseok Roh, Hwanhee Jung, Giljoo Nam, Jinseop Yeom, Hyunje Park, Sang Ho Yoon, Sangpil Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20644-20653
While recent 3D instance segmentation approaches show promising results based on transformer architectures they often fail to correctly identify instances with similar appearances. They also ambiguously determine edges leading to multiple misclassifications of adjacent edge points. In this work we introduce a novel framework called EASE to overcome these challenges and improve the perception of complex 3D instances. We first propose a semantic guidance network to leverage rich semantic knowledge from a language model as intelligent priors enhancing the functional understanding of real-world instances beyond relying solely on geometrical information. We explicitly instruct the basic instance queries using text embeddings of each instance to learn deep semantic details. Further we utilize the edge prediction module encouraging the segmentation network to be edge-aware. We extract voxel-wise edge maps from point features and use them as auxiliary information for learning edge cues. In our extensive experiments on large-scale benchmarks ScanNetV2 ScanNet200 S3DIS and STPLS3D our EASE outperforms existing state-of-the-art models demonstrating its superior performance.
****
Coherence As Texture - Passive Textureless 3D Reconstruction by Self-interference
Wei-Yu Chen, Aswin C. Sankaranarayanan, Anat Levin, Matthew O'Toole; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25058-25066
Passive depth estimation based on stereo or defocus relies on the presence of the texture on an object to resolve its depth. Hence recovering the depth of a textureless object-- for example a large white wall--is not just hard but perhaps even impossible. Or is it? We show that spatial coherence a property of natural light sources can be used to resolve the depth of a scene point even when it is textureless. Our approach relies on the idea that natural light scattered off a scene point is locally coherent with itself while incoherent with the light scattered from other surface points; we use this insight to design an optical setup that uses self-interference as a texture feature for estimating depth. Our lab prototype is capable of resolving the depths of textureless objects in sunlight as well as indoor lights.
****
Enhancing the Power of OOD Detection via Sample-Aware Model Selection
Feng Xue, Zi He, Yuan Zhang, Chuanlong Xie, Zhenguo Li, Falong Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17148-17157
In this work we present a novel perspective on detecting out-of-distribution (OOD) samples and propose an algorithm for sample-aware model selection to enhance

the effectiveness of OOD detection. Our algorithm determines for each test input
 which pre-trained models in the model zoo are capable of identifying the test i
nput as an OOD sample. If no such models exist in the model zoo the test input i
s classified as an in-distribution (ID) sample. We theoretically demonstrate tha
t our method maintains the true positive rate of ID samples and accurately ident
ifies OOD samples with high probability when there are a sufficient number of di
verse pre-trained models in the model zoo. Extensive experiments were conducted
to validate our method demonstrating that it leverages the complementarity among
 single-model detectors to consistently improve the effectiveness of OOD sample
identification. Compared to base-line methods our approach improved the relative
 performance by 65.40% and 37.25% on the CIFAR10 and ImageNet benchmarks respect
ively.
****
Collaborative Semantic Occupancy Prediction with Hybrid Feature Fusion in Connec
ted Automated Vehicles
Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andrea
s Festag, Alois Knoll; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2024, pp. 17996-18006
Collaborative perception in automated vehicles leverages the exchange of informa
tion between agents aiming to elevate perception results. Previous camera-based
collaborative 3D perception methods typically employ 3D bounding boxes or bird's
 eye views as representations of the environment. However these approaches fall
short in offering a comprehensive 3D environmental prediction. To bridge this ga
p we introduce the first method for collaborative 3D semantic occupancy predicti
on. Particularly it improves local 3D semantic occupancy predictions by hybrid f
usion of (i) semantic and occupancy task features and (ii) compressed orthogonal
 attention features shared between vehicles. Additionally due to the lack of a c
ollaborative perception dataset designed for semantic occupancy prediction we au
gment a current collaborative perception dataset to include 3D collaborative sem
antic occupancy labels for a more robust evaluation. The experimental findings h
ighlight that: (i) our collaborative semantic occupancy predictions excel above
the results from single vehicles by over 30% and (ii) models anchored on semanti
c occupancy outpace state-of-the-art collaborative 3D detection techniques in su
bsequent perception applications showcasing enhanced accuracy and enriched seman
tic-awareness in road environments.
****
Generative Multi-modal Models are Good Class Incremental Learners
Xusheng Cao, Haori Lu, Linlan Huang, Xialei Liu, Ming-Ming Cheng; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 28706-28717
In class incremental learning (CIL) scenarios the phenomenon of catastrophic for
getting caused by the classifier's bias towards the current task has long posed
a significant challenge. It is mainly caused by the characteristic of discrimina
tive models. With the growing popularity of the generative multi-modal models we
 would explore replacing discriminative models with generative ones for CIL. How
ever transitioning from discriminative to generative models requires addressing
two key challenges. The primary challenge lies in transferring the generated tex
tual information into the classification of distinct categories. Additionally it
 requires formulating the task of CIL within a generative framework. To this end
 we propose a novel generative multi-modal model (GMM) framework for class incre
mental learning. Our approach directly generates labels for images using an adap
ted generative model. After obtaining the detailed text we use a text encoder to
 extract text features and employ feature matching to determine the most similar
 label as the classification prediction. In the conventional CIL settings we ach
ieve significantly better results in long-sequence task scenarios. Under the Few
-shot CIL setting we have improved by at least 14% over the current state-of-the
-art methods with significantly less forgetting.
****
Low-Resource Vision Challenges for Foundation Models
Yunhua Zhang, Hazel Doughty, Cees G. M. Snoek; Proceedings of the IEEE/CVF Confe

rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21956-21966
Low-resource settings are well-established in natural lan- guage processing where many languages lack sufficient data for deep learning at scale. However low-resource problems are under-explored in computer vision. In this paper we address this gap and explore the challenges of low-resource image tasks with vision foundation models. We first collect a benchmark of genuinely low-resource image data covering historic maps circuit diagrams and mechanical drawings. These low-resource settings all share three challenges: data scarcity fine-grained differences and the distribution shift from natural images to the specialized domain of interest. While existing foundation models have shown impressive generalizability we find they cannot transfer well to our low-resource tasks. To begin to tackle the challenges of low-resource vision we introduce one simple baseline per challenge. Specifically we i) enlarge the data space by generative models ii) adopt the best sub-kernels to encode local regions for fine-grained difference discovery and iii) learn attention for specialized domains. Experiments on our three low-resource tasks demonstrate our proposals already provide a better baseline than transfer learning data aug- mentation and fine-grained methods. This highlights the unique characteristics and challenges of low-resource vision for foundation models that warrant further investigation. Project page: https://xiaobai1217.github.io/ Low-Resource-Vision/.
****
RGBD Objects in the Wild: Scaling Real-World 3D Object Learning from RGB-D Videos
Hongchi Xia, Yang Fu, Sifei Liu, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22378-22389
We introduce a new RGB-D object dataset captured in the wild called WildRGB-D. Unlike most existing real-world object-centric datasets which only come with RGB capturing the direct capture of the depth channel allows better 3D annotations and broader downstream applications. WildRGB-D comprises large-scale category-level RGB-D object videos which are taken using an iPhone to go around the objects in 360 degrees. It contains around 8500 recorded objects and nearly 20000 RGB-D videos across 46 common object categories. These videos are taken with diverse cluttered backgrounds with three setups to cover as many real-world scenarios as possible: (i) a single object in one video; (ii) multiple objects in one video; and (iii) an object with a static hand in one video. The dataset is annotated with object masks real-world scale camera poses and reconstructed aggregated point clouds from RGBD videos. We benchmark four tasks with WildRGB-D including novel view synthesis camera pose estimation object 6d pose estimation and object surface reconstruction. Our experiments show that the large-scale capture of RGB-D objects provides a large potential to advance 3D object learning. Our project page is https://wildrgbd.github.io/.
****
Shadow-Enlightened Image Outpainting
Hang Yu, Ruilin Li, Shaorong Xie, Jiayan Qiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7850-7860
Conventional image outpainting methods usually treat unobserved areas as unknown and extend the scene only in terms of semantic consistency thus overlooking the hidden information in shadows cast by unobserved areas such as the invisible shapes and semantics. In this paper we propose to extract and utilize the hidden information of unobserved areas from their shadows to enhance image outpainting. To this end we propose an end-to-end deep approach that explicitly looks into the shadows within the image. Specifically we extract shadows from the input image and identify instance-level shadow regions cast by the unobserved areas. Then the instance-level shadow representations are concatenated to predict the scene layout of each unobserved instance and outpaint the unobserved areas. Finally two discriminators are implemented to enhance alignment between the extended semantics and their shadows. In the experiments we show that our proposed approach provides complementary cues for outpainting and achieves considerable improvement on all datasets by adopting our approach as a plug-in module.
****

Towards Generalizable Tumor Synthesis
Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, Zongwei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11147-11158
Tumor synthesis enables the creation of artificial tumors in medical images facilitating the training of AI models for tumor detection and segmentation. However success in tumor synthesis hinges on creating visually realistic tumors that are generalizable across multiple organs and furthermore the resulting AI models being capable of detecting real tumors in images sourced from different domains (e.g. hospitals). This paper made a progressive stride toward generalizable tumor synthesis by leveraging a critical observation: early-stage tumors (< 2cm) tend to have similar imaging characteristics in computed tomography (CT) whether they originate in the liver pancreas or kidneys. We have ascertained that generative AI models e.g. Diffusion Models can create realistic tumors generalized to a range of organs even when trained on a limited number of tumor examples from only one organ. Moreover we have shown that AI models trained on these synthetic tumors can be generalized to detect and segment real tumors from CT volumes encompassing a broad spectrum of patient demographics imaging protocols and healthcare facilities.
****
Low-Res Leads the Way: Improving Generalization for Super-Resolution by Self-Supervised Learning
Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Haoze Sun, Xueyi Zou, Zhensong Zhang, Youliang Yan, Lei Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25857-25867
For image super-resolution (SR) bridging the gap between the performance on synthetic datasets and real-world degradation scenarios remains a challenge. This work introduces a novel "Low-Res Leads the Way" (LWay) training framework merging Supervised Pre-training with Self-supervised Learning to enhance the adaptability of SR models to real-world images. Our approach utilizes a low-resolution (LR) reconstruction network to extract degradation embeddings from LR images merging them with super-resolved outputs for LR reconstruction. Leveraging unseen LR images for self-supervised learning guides the model to adapt its modeling space to the target domain facilitating fine-tuning of SR models without requiring paired high-resolution (HR) images. The integration of Discrete Wavelet Transform (DWT) further refines the focus on high-frequency details. Extensive evaluations show that our method significantly improves the generalization and detail restoration capabilities of SR models on unseen real-world datasets outperforming existing methods. Our training regime is universally compatible requiring no network architecture modifications making it a practical solution for real-world SR applications.
****
BOTH2Hands: Inferring 3D Hands from Both Text Prompts and Body Dynamics
Wenqian Zhang, Molin Huang, Yuxuan Zhou, Juze Zhang, Jingyi Yu, Jingya Wang, Lan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2393-2404
The recently emerging text-to-motion advances have spired numerous attempts for convenient and interactive human motion generation. Yet existing methods are largely limited to generating body motions only without considering the rich two-hand motions let alone handling various conditions like body dynamics or texts. To break the data bottleneck we propose BOTH57M a novel multi-modal dataset for two-hand motion generation. Our dataset includes accurate motion tracking for the human body and hands and provides pair-wised finger-level hand annotations and body descriptions. We further provide a strong baseline method BOTH2Hands for the novel task: generating vivid two-hand motions from both implicit body dynamics and explicit text prompts. We first warm up two parallel body-to-hand and text-to-hand diffusion models and then utilize the cross-attention transformer for motion blending. Extensive experiments and cross-validations demonstrate the effectiveness of our approach and dataset for generating convincing two-hand motions from the hybrid body-and-textual conditions. Our dataset and code will be dissemi

nated to the community for future research which can be found at https://github.com/Godheritage/BOTH2Hands.
****
EpiDiff: Enhancing Multi-View Synthesis via Localized Epipolar-Constrained Diffusion
Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, Lu Sheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9784-9794
Generating multiview images from a single view facilitates the rapid generation of a 3D mesh conditioned on a single image. Recent methods that introduce 3D global representation into diffusion models have shown the potential to generate consistent multiviews but they have reduced generation speed and face challenges in maintaining generalizability and quality. To address this issue we propose EpiDiff a localized interactive multiview diffusion model. At the core of the proposed approach is to insert a lightweight epipolar attention block into the frozen diffusion model leveraging epipolar constraints to enable cross-view interaction among feature maps of neighboring views. The newly initialized 3D modeling module preserves the original feature distribution of the diffusion model exhibiting compatibility with a variety of base diffusion models. Experiments show that EpiDiff generates 16 multiview images in just 12 seconds and it surpasses previous methods in quality evaluation metrics including PSNR SSIM and LPIPS. Additionally EpiDiff can generate a more diverse distribution of views improving the reconstruction quality from generated multiviews. Please see the project page at https://huanngzh.github.io/EpiDiff/.
****
On the Faithfulness of Vision Transformer Explanations
Junyi Wu, Weitai Kang, Hao Tang, Yuan Hong, Yan Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10936-10945
To interpret Vision Transformers post-hoc explanations assign salience scores to input pixels providing human-understandable heatmaps. However whether these interpretations reflect true rationales behind the model's output is still underexplored. To address this gap we study the faithfulness criterion of explanations: the assigned salience scores should represent the influence of the corresponding input pixels on the model's predictions. To evaluate faithfulness we introduce Salience-guided Faithfulness Coefficient (SaCo) a novel evaluation metric leveraging essential information of salience distribution. Specifically we conduct pair-wise comparisons among distinct pixel groups and then aggregate the differences in their salience scores resulting in a coefficient that indicates the explanation's degree of faithfulness. Our explorations reveal that current metrics struggle to differentiate between advanced explanation methods and Random Attribution thereby failing to capture the faithfulness property. In contrast our proposed SaCo offers a reliable faithfulness measurement establishing a robust metric for interpretations. Furthermore our SaCo demonstrates that the use of gradient and multi-layer aggregation can markedly enhance the faithfulness of attention-based explanation shedding light on potential paths for advancing Vision Transformer explainability.
****
Pixel-level Semantic Correspondence through Layout-aware Representation Learning and Multi-scale Matching Integration
Yixuan Sun, Zhangyue Yin, Haibo Wang, Yan Wang, Xipeng Qiu, Weifeng Ge, Wenqiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17047-17056
Establishing precise semantic correspondence across object instances in different images is a fundamental and challenging task in computer vision. In this task difficulty arises often due to three challenges: confusing regions with similar appearance inconsistent object scale and indistinguishable nearby pixels. Recognizing these challenges our paper proposes a novel semantic matching pipeline named LPMFlow toward extracting fine-grained semantics and geometry layouts for building pixel-level semantic correspondences. LPMFlow consists of three modules ea

ch addressing one of the aforementioned challenges. The layout-aware representation learning module uniformly encodes source and target tokens to distinguish pixels or regions with similar appearances but different geometry semantics. The progressive feature superresolution module outputs four sets of 4D correlation tensors to generate accurate semantic flow between objects in different scales. Finally the matching flow integration and refinement module is exploited to fuse matching flow in different scales to give the final flow predictions. The whole pipeline can be trained end-to-end with a balance of computational cost and correspondence details. Extensive experiments based on benchmarks such as SPair-71K PF-PASCAL and PF-WILLOW have proved that the proposed method can well tackle the three challenges and outperform the previous methods especially in more stringent settings. Code is available at https://github.com/YXSUNMADMAX/LPMFlow.
****
Learning Spatial Features from Audio-Visual Correspondence in Egocentric Videos
Sagnik Majumder, Ziad Al-Halah, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27058-27068
We propose a self-supervised method for learning representations based on spatial audio-visual correspondences in egocentric videos. Our method uses a masked auto-encoding framework to synthesize masked binaural audio through the synergy of audio and vision thereby learning useful spatial relationships between the two modalities. We use our pretrained features to tackle two downstream video tasks requiring spatial understanding in social scenarios: active speaker detection and spatial audio denoising. Through extensive experiments we show that our features are generic enough to improve over multiple state-of-the-art baselines on both tasks on two challenging egocentric video datasets that offer binaural audio EgoCom and EasyCom. Project: http://vision.cs.utexas.edu/ projects/ego_av_corr.
****
DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models
Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, Kwan-Yee K. Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 958-968
We present DreamAvatar a text-and-shape guided framework for generating high-quality 3D human avatars with controllable poses. While encouraging results have been reported by recent methods on text-guided 3D common object generation generating high-quality human avatars remains an open challenge due to the complexity of the human body's shape pose and appearance. We propose DreamAvatar to tackle this challenge which utilizes a trainable NeRF for predicting density and color for 3D points and pretrained text-to-image diffusion models for providing 2D self-supervision. Specifically we leverage the SMPL model to provide shape and pose guidance for the generation. We introduce a dual-observation-space design that involves the joint optimization of a canonical space and a posed space that are related by a learnable deformation field. This facilitates the generation of more complete textures and geometry faithful to the target pose. We also jointly optimize the losses computed from the full body and from the zoomed-in 3D head to alleviate the common multi-face "Janus" problem and improve facial details in the generated avatars. Extensive evaluations demonstrate that DreamAvatar significantly outperforms existing methods establishing a new state-of-the-art for text-and-shape guided 3D human avatar generation.
****
Dynamic Graph Representation with Knowledge-aware Attention for Histopathology Whole Slide Image Analysis
Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, Yonghong He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11323-11332
Histopathological whole slide images (WSIs) classification has become a foundation task in medical microscopic imaging processing. Prevailing approaches involve learning WSIs as instance-bag representations emphasizing significant instances but struggling to capture the interactions between instances. Additionally conventional graph representation methods utilize explicit spatial positions to cons

truct topological structures but restrict the flexible interaction capabilities between instances at arbitrary locations particularly when spatially distant. In response we propose a novel dynamic graph representation algorithm that conceptualizes WSIs as a form of the knowledge graph structure. Specifically we dynamically construct neighbors and directed edge embeddings based on the head and tail relationships between instances. Then we devise a knowledge-aware attention mechanism that can update the head node features by learning the joint attention score of each neighbor and edge. Finally we obtain a graph-level embedding through the global pooling process of the updated head serving as an implicit representation for the WSI classification. Our end-to-end graph representation learning approach has outperformed the state-of-the-art WSI analysis methods on three TCGA benchmark datasets and in-house test sets. Our code is available at https://github.com/WonderLandxD/WiKG.
****
Brain Decodes Deep Nets
Huzheng Yang, James Gee, Jianbo Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23030-23040
We developed a tool for visualizing and analyzing large pre-trained vision models by mapping them onto the brain thus exposing their hidden inside. Our innovation arises from a surprising usage of brain encoding: predicting brain fMRI measurements in response to images. We report two findings. First explicit mapping between the brain and deep-network features across dimensions of space layers scales and channels is crucial. This mapping method FactorTopy is plug-and-play for any deep-network; with it one can paint a picture of the network onto the brain (literally!). Second our visualization shows how different training methods matter: they lead to remarkable differences in hierarchical organization and scaling behavior growing with more data or network capacity. It also provides insight into fine-tuning: how pre-trained models change when adapting to small datasets. We found brain-like hierarchically organized network suffer less from catastrophic forgetting after fine-tuned.
****
Semantics Distortion and Style Matter: Towards Source-free UDA for Panoramic Segmentation
Xu Zheng, Pengyuan Zhou, Athanasios V. Vasilakos, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27885-27895
This paper addresses an interesting yet challenging problem-- source-free unsupervised domain adaptation (SFUDA) for pinhole-to-panoramic semantic segmentation--given only a pinhole image-trained model (i.e. source) and unlabeled panoramic images (i.e. target). Tackling this problem is nontrivial due to the semantic mismatches style discrepancies and inevitable distortion of panoramic images. To this end we propose a novel method that utilizes Tangent Projection (TP) as it has less distortion and meanwhile slits the equirectangular projection (ERP) with a fixed FoV to mimic the pinhole images. Both projections are shown effective in extracting knowledge from the source model. However the distinct projection discrepancies between source and target domains impede the direct knowledge transfer; thus we propose a panoramic prototype adaptation module (PPAM) to integrate panoramic prototypes from the extracted knowledge for adaptation. We then impose the loss constraints on both predictions and prototypes and propose a cross-dual attention module (CDAM) at the feature level to better align the spatial and channel characteristics across the domains and projections. Both knowledge extraction and transfer processes are synchronously updated to reach the best performance. Extensive experiments on the synthetic and real-world benchmarks including outdoor and indoor scenarios demonstrate that our method achieves significantly better performance than prior SFUDA methods for pinhole-to-panoramic adaptation.
****
Bidirectional Autoregessive Diffusion Model for Dance Generation
Canyu Zhang, Youbao Tang, Ning Zhang, Ruei-Sung Lin, Mei Han, Jing Xiao, Song Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 687-696

Dance serves as a powerful medium for expressing human emotions but the lifelike generation of dance is still a considerable challenge. Recently diffusion models have showcased remarkable generative abilities across various domains. They hold promise for human motion generation due to their adaptable many-to-many nature. Nonetheless current diffusion-based motion generation models often create entire motion sequences directly and unidirectionally lacking focus on the motion with local and bidirectional enhancement. When choreographing high-quality dance movements people need to take into account not only the musical context but also the nearby music-aligned dance motions. To authentically capture human behavior we propose a Bidirectional Autoregressive Diffusion Model (BADM) for music-to-dance generation where a bidirectional encoder is built to enforce that the generated dance is harmonious in both the forward and backward directions. To make the generated dance motion smoother a local information decoder is built for local motion enhancement. The proposed framework is able to generate new motions based on the input conditions and nearby motions which foresees individual motion slices iteratively and consolidates all predictions. To further refine the synchronicity between the generated dance and the beat the beat information is incorporated as an input to generate better music-aligned dance movements. Experimental results demonstrate that the proposed model achieves state-of-the-art performance compared to existing unidirectional approaches on the prominent benchmark for music-to-dance generation.

****

Align Before Adapt: Leveraging Entity-to-Region Alignments for Generalizable Video Action Recognition

Yifei Chen, Dapeng Chen, Ruijin Liu, Sai Zhou, Wenyuan Xue, Wei Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18688-18698

Large-scale visual-language pre-trained models have achieved significant success in various video tasks. However most existing methods follow an "adapt then align" paradigm which adapts pre-trained image encoders to model video-level representations and utilizes one-hot or text embedding of the action labels for supervision. This paradigm overlooks the challenge of mapping from static images to complicated activity concepts. In this paper we propose a novel "Align before Adapt" (ALT) paradigm. Prior to adapting to video representation learning we exploit the entity-to-region alignments for each frame. The alignments are fulfilled by matching the region-aware image embeddings to an offline-constructed text corpus. With the aligned entities we feed their text embeddings to a transformer-based video adapter as the queries which can help extract the semantics of the most important entities from a video to a vector. This paradigm reuses the visual-language alignment of VLP during adaptation and tries to explain an action by the underlying entities. This helps understand actions by bridging the gap with complex activity semantics particularly when facing unfamiliar or unseen categories. ALT demonstrates competitive performance while maintaining remarkably low computational costs. In fully supervised experiments it achieves 88.1% top-1 accuracy on Kinetics-400 with only 4947 GFLOPs. Moreover ALT outperforms the previous state-of-the-art methods in both zero-shot and few-shot experiments emphasizing its superior generalizability across various learning scenarios.

****

GOV-NeSF: Generalizable Open-Vocabulary Neural Semantic Fields

Yunsong Wang, Hanlin Chen, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20443-20453

Recent advancements in vision-language foundation models have significantly enhanced open-vocabulary 3D scene understanding. However the generalizability of existing methods is constrained due to their framework designs and their reliance on 3D data. We address this limitation by introducing Generalizable Open-Vocabulary Neural Semantic Fields (GOV-NeSF) a novel approach offering a generalizable implicit representation of 3D scenes with open-vocabulary semantics. We aggregate the geometry-aware features using a cost volume and propose a Multi-view Joint Fusion module to aggregate multi-view features through a cross-view attention mechanism which effectively predicts view-specific blending weights for both color

s and open-vocabulary features. Remarkably our GOV-NeSF exhibits state-of-the-ar
t performance in both 2D and 3D open-vocabulary semantic segmentation eliminatin
g the need for ground truth semantic labels or depth priors and effectively gene
ralize across scenes and datasets without fine-tuning.
****
FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation
Shuai Yang, Yifan Zhou, Ziwei Liu, Chen Change Loy; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8703-871
2
The remarkable efficacy of text-to-image diffusion models has motivated extensiv
e exploration of their potential application in video domains. Zero-shot methods
 seek to extend image diffusion models to videos without necessitating model tra
ining. Recent methods mainly focus on incorporating inter-frame correspondence i
nto attention mechanisms. However the soft constraint imposed on determining whe
re to attend to valid features can sometimes be insufficient resulting in tempor
al inconsistency. In this paper we introduce FRESCO intra-frame correspondence a
longside inter-frame correspondence to establish a more robust spatial-temporal
constraint. This enhancement ensures a more consistent transformation of semanti
cally similar content across frames. Beyond mere attention guidance our approach
 involves an explicit update of features to achieve high spatial-temporal consis
tency with the input video significantly improving the visual coherence of the r
esulting translated videos. Extensive experiments demonstrate the effectiveness
of our proposed framework in producing high-quality coherent videos marking a no
table improvement over existing zero-shot methods.
****
Dual-Scale Transformer for Large-Scale Single-Pixel Imaging
Gang Qu, Ping Wang, Xin Yuan; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2024, pp. 25327-25337
Single-pixel imaging (SPI) is a potential computational imaging technique which
produces image by solving an ill-posed reconstruction problem from few measureme
nts captured by a single-pixel detector. Deep learning has achieved impressive s
uccess on SPI reconstruction. However previous poor reconstruction performance a
nd impractical imaging model limit its real-world applications. In this paper we
 propose a deep unfolding network with hybrid-attention Transformer on Kronecker
 SPI model dubbed HATNet to improve the imaging quality of real SPI cameras. Spe
cifically we unfold the computation graph of the iterative shrinkage-thresholdin
g algorithm (ISTA) into two alternative modules: efficient tensor gradient desce
nt and hybrid-attention multi-scale denoising. By virtue of Kronecker SPI the gr
adient descent module can avoid high computational overheads rooted in previous
gradient descent modules based on vectorized SPI. The denoising module is an enc
oder-decoder architecture powered by dual-scale spatial attention for high- and
low-frequency aggregation and channel attention for global information recalibra
tion. Moreover we build a SPI prototype to verify the effectiveness of the propo
sed method. Extensive experiments on synthetic and real data demonstrate that ou
r method achieves the state-of-the-art performance. The source code and pre-trai
ned models are available at https://github.com/Gang-Qu/HATNet-SPI.
****
Towards Robust 3D Object Detection with LiDAR and 4D Radar Fusion in Various Wea
ther Conditions
Yujeong Chae, Hyeonseong Kim, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15162-15172
Detecting objects in 3D under various (normal and adverse) weather conditions is
 essential for safe autonomous driving systems. Recent approaches have focused o
n employing weather-insensitive 4D radar sensors and leveraging them with other
modalities such as LiDAR. However they fuse multi-modal information without cons
idering the sensor characteristics and weather conditions and lose some height i
nformation which could be useful for localizing 3D objects. In this paper we pro
pose a novel framework for robust LiDAR and 4D radar-based 3D object detection.
Specifically we propose a 3D-LRF module that considers the distinct patterns the
y exhibit in 3D space (e.g. precise 3D mapping of LiDAR and wide-range weather-i

nsensitive measurement of 4D radar) and extract fusion features based on their 3D spatial relationship. Then our weather-conditional radar-flow gating network modulates the information flow of fusion features depending on weather conditions and obtains enhanced feature that effectively incorporates the strength of two domains under various weather conditions. The extensive experiments demonstrate that our model achieves SoTA performance for 3D object detection under various weather conditions.
****

Enhancing 3D Fidelity of Text-to-3D using Cross-View Correspondences
Seungwook Kim, Kejie Li, Xueqing Deng, Yichun Shi, Minsu Cho, Peng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10649-10658
Leveraging multi-view diffusion models as priors for 3D optimization have alleviated the problem of 3D consistency e.g. the Janus face problem or the content drift problem in zero-shot text-to-3D models. However the 3D geometric fidelity of the output remains an unresolved issue; albeit the rendered 2D views are realistic the underlying geometry may contain errors such as unreasonable concavities. In this work we propose CorrespondentDream an effective method to leverage annotation-free cross-view correspondences yielded from the diffusion U-Net to provide additional 3D prior to the NeRF optimization process. We find that these correspondences are strongly consistent with human perception and by adopting it in our loss design we are able to produce NeRF models with geometries that are more coherent with common sense e.g. more smoothed object surface yielding higher 3D fidelity. We demonstrate the efficacy of our approach through various comparative qualitative results and a solid user study.
****

Bezier Everywhere All at Once: Learning Drivable Lanes as Bezier Graphs
Hugh Blayney, Hanlin Tian, Hamish Scott, Nils Goldbeck, Chess Stetson, Panagiotis Angeloudis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15365-15374
Knowledge of lane topology is a core problem in autonomous driving. Aerial imagery can provide high resolution quickly updatable lane source data but detecting lanes from such data has so far been an expensive manual process or where automated solutions exist undrivable and requiring of downstream processing. We propose a method for large-scale lane topology extraction from aerial imagery while ensuring that the resulting lanes are realistic and drivable by introducing a novel Bezier Graph shared parameterisation of Bezier curves. We develop a transformer-based model to predict these Bezier Graphs from input aerial images demonstrating competitive results on the UrbanLaneGraph dataset. We demonstrate that our method generates realistic lane graphs which require both minimal input and minimal downstream processing. We make our code publicly available at https://github.com/driskai/BGFormer
****

SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting
Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, Zeyu Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1606-1616
We present SplattingAvatar a hybrid 3D representation of photorealistic human avatars with Gaussian Splatting embedded on a triangle mesh which renders over 300 FPS on a modern GPU and 30 FPS on a mobile device. We disentangle the motion and appearance of a virtual human with explicit mesh geometry and implicit appearance modeling with Gaussian Splatting. The Gaussians are defined by barycentric coordinates and displacement on a triangle mesh as Phong surfaces. We extend lifted optimization to simultaneously optimize the parameters of the Gaussians while walking on the triangle mesh. SplattingAvatar is a hybrid representation of virtual humans where the mesh represents low-frequency motion and surface deformation while the Gaussians take over the high-frequency geometry and detailed appearance. Unlike existing deformation methods that rely on an MLP-based linear blend skinning (LBS) field for motion we control the rotation and translation of the

Gaussians directly by mesh which empowers its compatibility with various animati
on techniques e.g. skeletal animation blend shapes and mesh editing. Trainable f
rom monocular videos for both full-body and head avatars SplattingAvatar shows s
tate-of-the-art rendering quality across multiple datasets.
****
MoSAR: Monocular Semi-Supervised Model for Avatar Reconstruction using Different
iable Shading
Abdallah Dib, Luiz Gustavo Hafemann, Emeline Got, Trevor Anderson, Amin Fadaeine
jad, Rafael M. O. Cruz, Marc-André Carbonneau; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1770-1780
Reconstructing an avatar from a portrait image has many applications in multimed
ia but remains a challenging research problem. Extracting reflectance maps and g
eometry from one image is ill-posed: recovering geometry is a one-to-many mappin
g problem and reflectance and light are difficult to disentangle. Accurate geome
try and reflectance can be captured under the controlled conditions of a light s
tage but it is costly to acquire large datasets in this fashion. Moreover traini
ng solely with this type of data leads to poor generalization with in-the-wild i
mages. This motivates the introduction of MoSAR a method for 3D avatar generatio
n from monocular images. We propose a semi-supervised training scheme that impro
ves generalization by learning from both light stage and in-the-wild datasets. T
his is achieved using a novel differentiable shading formulation. We show that o
ur approach effectively disentangles the intrinsic face parameters producing rel
ightable avatars. As a result MoSAR estimates a richer set of skin reflectance m
aps and generates more realistic avatars than existing state-of-the-art methods.
 We also release a new dataset that provides intrinsic face attributes (diffuse
specular ambient occlusion and translucency maps) for 10k subjects.
****
Bridging Remote Sensors with Multisensor Geospatial Foundation Models
Boran Han, Shuai Zhang, Xingjian Shi, Markus Reichstein; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 278
52-27862
In the realm of geospatial analysis the diversity of remote sensors encompassing
 both optical and microwave technologies offers a wealth of distinct observation
al capabilities. Recognizing this we present msGFM a multisensor geospatial foun
dation model that effectively unifies data from four key sensor modalities. This
 integration spans an expansive dataset of two million multisensor images. msGFM
 is uniquely adept at handling both paired and unpaired sensor data. For data or
iginating from identical geolocations our model employs an innovative cross-sens
or pretraining approach in masked image modeling enabling the synthesis of joint
 representations from diverse sensors. msGFM incorporating four remote sensors u
pholds strong performance forming a comprehensive model adaptable to various sen
sor types. msGFM has demonstrated enhanced proficiency in a range of both single
-sensor and multisensor downstream tasks. These include scene classification seg
mentation cloud removal and pan-sharpening. A key discovery of our research is t
hat representations derived from natural images are not always compatible with t
he distinct characteristics of geospatial remote sensors underscoring the limita
tions of existing representations in this field. Our work can serve as a guide f
or developing multisensor geospatial pretraining models paving the way for more
advanced geospatial capabilities. Code can be found at \url https://github.com/b
oranhan/Geospatial_Foundation_Models
****
Can I Trust Your Answer? Visually Grounded Video Question Answering
Junbin Xiao, Angela Yao, Yicong Li, Tat-Seng Chua; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13204-132
14
We study visually grounded VideoQA in response to the emerging trends of utilizi
ng pretraining techniques for video- language understanding. Specifically by for
cing vision- language models (VLMs) to answer questions and simultane- ously pro
vide visual evidence we seek to ascertain the extent to which the predictions of
 such techniques are genuinely anchored in relevant video content versus spuriou

s corre- lations from language or irrelevant visual context. Towards this we con struct NExT-GQA - an extension of NExT-QA with 10.5K temporal grounding (or loca tion) labels tied to the original QA pairs. With NExT-GQA we scrutinize a series of state-of-the-art VLMs. Through post-hoc atten- tion analysis we find that th ese models are extremely weak in substantiating the answers despite their strong QA per- formance. This exposes the limitation of current VLMs in making reliabl e predictions. As a remedy we further explore and propose a grounded-QA method v ia Gaussian mask optimization and cross-modal learning. Experiments with differe nt backbones demonstrate that this grounding mechanism improves both grounding a nd QA. With these efforts we aim to push towards trustworthy VLMs in VQA systems . Our dataset and code are available at https://github.com/doc-doc/NExT-GQA.
****
RankED: Addressing Imbalance and Uncertainty in Edge Detection Using Ranking-bas ed Losses
Bedrettin Cetinkaya, Sinan Kalkan, Emre Akbas; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3239-3249
Detecting edges in images suffers from the problems of (P1) heavy imbalance betw een positive and negative classes as well as (P2) label uncertainty owing to dis agreement between different annotators. Existing solutions address P1 using clas s-balanced cross-entropy loss and dice loss and P2 by only predicting edges agre ed upon by most annotators. In this paper we propose RankED a unified ranking-ba sed approach that addresses both the imbalance problem (P1) and the uncertainty problem (P2). RankED tackles these two problems with two components: One compone nt which ranks positive pixels over negative pixels and the second which promote s high confidence edge pixels to have more label certainty. We show that RankED outperforms previous studies and sets a new state-of-the-art on NYUD-v2 BSDS500 and Multi-cue datasets. Code is available at https://ranked-cvpr24.github.io.
****
DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans
Akash Sengupta, Thiemo Alldieck, Nikos Kolotouros, Enric Corona, Andrei Zanfir, Cristian Sminchisescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1439-1449
We present DiffHuman a probabilistic method for photorealistic 3D human reconstr uction from a single RGB image. Despite the ill-posed nature of this problem mos t methods are deterministic and output a single solution often resulting in a la ck of geometric detail and blurriness in unseen or uncertain regions. In contras t DiffHuman predicts a probability distribution over 3D reconstructions conditio ned on an input 2D image which allows us to sample multiple detailed 3D avatars that are consistent with the image. DiffHuman is implemented as a conditional di ffusion model that denoises pixel-aligned 2D observations of an underlying 3D sh ape representation. During inference we may sample 3D avatars by iteratively den oising 2D renders of the predicted 3D representation. Furthermore we introduce a generator neural network that approximates rendering with considerably reduced runtime (55x speed up) resulting in a novel dual-branch diffusion framework. Our experiments show that DiffHuman can produce diverse and detailed reconstruction s for the parts of the person that are unseen or uncertain in the input image wh ile remaining competitive with the state-of-the-art when reconstructing visible surfaces.
****
SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution
Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, Lei Zhang; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2024, pp. 25456-25467
Owe to the powerful generative priors the pre-trained text-to-image (T2I) diffus ion models have become increasingly popular in solving the real-world image supe r-resolution problem. However as a consequence of the heavy quality degradation of input low-resolution (LR) images the destruction of local structures can lead to ambiguous image semantics. As a result the content of reproduced high-resolu tion image may have semantic errors deteriorating the super-resolution performan ce. To address this issue we present a semantics-aware approach to better preser

ve the semantic fidelity of generative real-world image super-resolution. First we train a degradation-aware prompt extractor which can generate accurate soft and hard semantic prompts even under strong degradation. The hard semantic prompts refer to the image tags aiming to enhance the local perception ability of the T2I model while the soft semantic prompts compensate for the hard ones to provide additional representation information. These semantic prompts encourage the T2I model to generate detailed and semantically accurate results. Furthermore during the inference process we integrate the LR images into the initial sampling noise to mitigate the diffusion model's tendency to generate excessive random details. The experiments show that our method can reproduce more realistic image details and hold better the semantics. The source code of our method can be found at https://github.com/cswry/SeeSR
****
Permutation Equivariance of Transformers and Its Applications
Hengyuan Xu, Liyao Xiang, Hangyu Ye, Dixi Yao, Pengzhi Chu, Baochun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5987-5996
Revolutionizing the field of deep learning Transformer-based models have achieved remarkable performance in many tasks. Recent research has recognized these models are robust to shuffling but are limited to inter-token permutation in the forward propagation. In this work we propose our definition of permutation equivariance a broader concept covering both inter- and intra- token permutation in the forward and backward propagation of neural networks. We rigorously proved that such permutation equivariance property can be satisfied on most vanilla Transformer-based models with almost no adaptation. We examine the property over a range of state-of-the-art models including ViT Bert GPT and others with experimental validations. Further as a proof-of-concept we explore how real-world applications including privacy-enhancing split learning and model authorization could exploit the permutation equivariance property which implicates wider intriguing application scenarios. The code is available at https://github.com/Doby-Xu/ST
****
Polos: Multimodal Metric Learning from Human Feedback for Image Captioning
Yuiga Wada, Kanta Kaneda, Daichi Saito, Komei Sugiura; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13559-13568
Establishing an automatic evaluation metric that closely aligns with human judgments is essential for effectively developing image captioning models. Recent data-driven metrics have demonstrated a stronger correlation with human judgments than classic metrics such as CIDEr; however they lack sufficient capabilities to handle hallucinations and generalize across diverse images and texts partially because they compute scalar similarities merely using embeddings learned from tasks unrelated to image captioning evaluation. In this study we propose Polos a supervised automatic evaluation metric for image captioning models. Polos computes scores from multimodal inputs using a parallel feature extraction mechanism that leverages embeddings trained through large-scale contrastive learning. To train Polos we introduce Multimodal Metric Learning from Human Feedback (M2LHF) a framework for developing metrics based on human feedback. We constructed the Polaris dataset which comprises 131K human judgments from 550 evaluators which is approximately ten times larger than standard datasets. Our approach achieved state-of-the-art performance on Composite Flickr8K-Expert Flickr8K-CF PASCAL-50S FOIL and the Polaris dataset thereby demonstrating its effectiveness and robustness.
****
Detours for Navigating Instructional Videos
Kumar Ashutosh, Zihui Xue, Tushar Nagarajan, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18804-18815
We introduce the video detours problem for navigating instructional videos. Given a source video and a natural language query asking to alter the how-to video's current path of execution in a certain way the goal is to find a related "detour video" that satisfies the requested alteration. To address this challenge we p

ropose VidDetours a novel video-language approach that learns to retrieve the targeted temporal segments from a large repository of how-to's using video-and-text conditioned queries. Furthermore we devise a language-based pipeline that exploits how-to video narration text to create weakly supervised training data. We demonstrate our idea applied to the domain of how-to cooking videos where a user can detour from their current recipe to find steps with alternate ingredients tools and techniques. Validating on a ground truth annotated dataset of 16K samples we show our model's significant improvements over best available methods for video retrieval and question answering with recall rates exceeding the state of the art by 35%.
****
Discontinuity-preserving Normal Integration with Auxiliary Edges
Hyomin Kim, Yucheol Jung, Seungyong Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11915-11923
Many surface reconstruction methods incorporate normal integration which is a process to obtain a depth map from surface gradients. In this process the input may represent a surface with discontinuities e.g. due to self-occlusion. To reconstruct an accurate depth map from the input normal map hidden surface gradients occurring from the jumps must be handled. To model these jumps correctly we design a novel discretization for the domain of normal integration. Our key idea is to introduce auxiliary edges which bridge between piecewise-smooth planes in the domain so that the magnitude of hidden jumps can be explicitly expressed on finite elements. Using the auxiliary edges we design a novel algorithm to optimize the discontinuity and the depth map from the input normal map. Our method optimizes discontinuities by using a combination of iterative re-weighted least squares and iterative filtering of the jump magnitudes on auxiliary edges to provide strong sparsity regularization. Compared to previous discontinuity-preserving normal integration methods which model the magnitude of jumps only implicitly our method reconstructs subtle discontinuities accurately thanks to our explicit representation allowing for strong sparsity regularization.
****
DrivingGaussian: Composite Gaussian Splatting for Surrounding Dynamic Autonomous Driving Scenes
Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21634-21643
We present DrivingGaussian an efficient and effective framework for surrounding dynamic autonomous driving scenes. For complex scenes with moving objects we first sequentially and progressively model the static background of the entire scene with incremental static 3D Gaussians. We then leverage a composite dynamic Gaussian graph to handle multiple moving objects individually reconstructing each object and restoring their accurate positions and occlusion relationships within the scene. We further use a LiDAR prior for Gaussian Splatting to reconstruct scenes with greater details and maintain panoramic consistency. DrivingGaussian outperforms existing methods in dynamic driving scene reconstruction and enables photorealistic surround-view synthesis with high-fidelity and multi-camera consistency. Our project page is at: https://github.com/VDIGPKU/DrivingGaussian.
****
Self-Supervised Multi-Object Tracking with Path Consistency
Zijia Lu, Bing Shuai, Yanbei Chen, Zhenlin Xu, Davide Modolo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19016-19026
In this paper we propose a novel concept of path consistency to learn robust object matching without using manual object identity supervision. Our key idea is that to track a object through frames we can obtain multiple different association results from a model by varying the frames it can observe i.e. skipping frames in observation. As the differences in observations do not alter the identities of objects the obtained association results should be consistent. Based on this rationale we generate multiple observation paths each specifying a different set of frames to be skipped and formulate the Path Consistency Loss that enforces t

he association results are consistent across different observation paths. We use the proposed loss to train our object matching model with only self-supervision. By extensive experiments on three tracking datasets (MOT17 PersonPath22 KITTI) we demonstrate that our method outperforms existing unsupervised methods with consistent margins on various evaluation metrics and even achieves performance close to supervised methods.
****

Unsupervised Keypoints from Pretrained Diffusion Models
Eric Hedlin, Gopal Sharma, Shweta Mahajan, Xingzhe He, Hossam Isack, Abhishek Kar, Helge Rhodin, Andrea Tagliasacchi, Kwang Moo Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22820-22830
Unsupervised learning of keypoints and landmarks has seen significant progress with the help of modern neural network architectures but performance is yet to match the supervised counterpart making their practicability questionable. We leverage the emergent knowledge within text-to-image diffusion models towards more robust unsupervised keypoints. Our core idea is to find text embeddings that would cause the generative model to consistently attend to compact regions in images (i.e. keypoints). To do so we simply optimize the text embedding such that the cross-attention maps within the denoising network are localized as Gaussians with small standard deviations. We validate our performance on multiple datasets: the CelebA CUB-200-2011 Tai-Chi-HD DeepFashion and Human3.6m datasets. We achieve significantly improved accuracy sometimes even outperforming supervised ones particularly for data that is non-aligned and less curated. Our code is publicly available at https://stablekeypoints.github.io/.
****

Resolution Limit of Single-Photon LiDAR
Stanley H. Chan, Hashan K. Weerasooriya, Weijian Zhang, Pamela Abshire, Istvan Gyongy, Robert K. Henderson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25307-25316
Single-photon Light Detection and Ranging (LiDAR) systems are often equipped with an array of detectors for improved spatial resolution and sensing speed. However given a fixed amount of flux produced by the laser transmitter across the scene the per-pixel Signal-to-Noise Ratio (SNR) will decrease when more pixels are packed in a unit space. This presents a fundamental trade-off between the spatial resolution of the sensor array and the SNR received at each pixel. Theoretical characterization of this fundamental limit is explored. By deriving the photon arrival statistics and introducing a series of new approximation techniques the Mean Squared Error (MSE) of the maximum-likelihood estimator of the time delay is derived. The theoretical predictions align well with simulations and real data.
****

Flatten Long-Range Loss Landscapes for Cross-Domain Few-Shot Learning
Yixiong Zou, Yicong Liu, Yiman Hu, Yuhua Li, Ruixuan Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23575-23584
Cross-domain few-shot learning (CDFSL) aims to acquire knowledge from limited training data in the target domain by leveraging prior knowledge transferred from source domains with abundant training samples. CDFSL faces challenges in transferring knowledge across dissimilar domains and fine-tuning models with limited training data. To address these challenges we initially extend the analysis of loss landscapes from the parameter space to the representation space which allows us to simultaneously interpret the transferring and fine-tuning difficulties of CDFSL models. We observe that sharp minima in the loss landscapes of the representation space result in representations that are hard to transfer and fine-tune. Moreover existing flatness-based methods have limited generalization ability due to their short-range flatness. To enhance the transferability and facilitate fine-tuning we introduce a simple yet effective approach to achieve long-range flattening of the minima in the loss landscape. This approach considers representations that are differently normalized as minima in the loss landscape and flatten

s the high-loss region in the middle by randomly sampling interpolated represent ations. We implement this method as a new normalization layer that replaces the original one in both CNNs and ViTs. This layer is simple and lightweight introdu cing only a minimal number of additional parameters. Experimental results on 8 d atasets demonstrate that our approach outperforms state-of-the-art methods in te rms of average accuracy. Moreover our method achieves performance improvements o f up to 9% compared to the current best approaches on individual datasets. Our c ode will be released.
****
Improving Distant 3D Object Detection Using 2D Box Supervision
Zetong Yang, Zhiding Yu, Chris Choy, Renhao Wang, Anima Anandkumar, Jose M. Alva rez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2024, pp. 14853-14863
Improving the detection of distant 3d objects is an important yet challenging ta sk. For camera-based 3D perception the annotation of 3d bounding relies heavily on LiDAR for accurate depth information. As such the distance of annotation is o ften limited due to the sparsity of LiDAR points on distant objects which hamper s the capability of existing detectors for long-range scenarios. We address this challenge by considering only 2D box supervision for distant objects since they are easy to annotate. We propose LR3D a framework that learns to recover the mi ssing depth of distant objects. LR3D adopts an implicit projection head to learn the generation of mapping between 2D boxes and depth using the 3D supervision o n close objects. This mapping allows the depth estimation of distant objects con ditioned on their 2D boxes making long-range 3D detection with 2D supervision fe asible. Experiments show that without distant 3D annotations LR3D allows camera- based methods to detect distant objects (over 200m) with comparable accuracy to full 3D supervision. Our framework is general and could widely benefit 3D detect ion methods to a large extent.
****
HDQMF: Holographic Feature Decomposition Using Quantum Algorithms
Prathyush Prasanth Poduval, Zhuowen Zou, Mohsen Imani; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10978 -10987
This paper addresses the decomposition of holographic feature vectors in Hyperdi mensional Computing (HDC) aka Vector Symbolic Architectures (VSA). HDC uses high -dimensional vectors with brain-like properties to represent symbolic informatio n and leverages efficient operators to construct and manipulate complexly struct ured data in a cognitive fashion. Existing models face challenges in decomposing these structures a process crucial for understanding and interpreting a composi te hypervector. We address this challenge by proposing the HDC Memorized-Factori zation Problem that captures the common patterns of construction in HDC models. To solve this problem efficiently we introduce HDQMF a HyperDimensional Quantum Memorized-Factorization algorithm. HDQMF is unique in its approach utilizing qua ntum computing to offer efficient solutions. It modifies crucial steps in Grover 's algorithm to achieve hypervector decomposition achieving quadratic speed-up.
****
Diffusion-based Blind Text Image Super-Resolution
Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, Liheng Bian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2024, pp. 25827-25836
Recovering degraded low-resolution text images is challenging especially for Chi nese text images with complex strokes and severe degradation in real-world scena rios. Ensuring both text fidelity and style realness is crucial for high-quality text image super-resolution. Recently diffusion models have achieved great succ ess in natural image synthesis and restoration due to their powerful data distri bution modeling abilities and data generation capabilities. In this work we prop ose an Image Diffusion Model (IDM) to restore text images with realistic styles. For diffusion models they are not only suitable for modeling realistic image di stribution but also appropriate for learning text distribution. Since text prior is important to guarantee the correctness of the restored text structure accord

ing to existing arts we also propose a Text Diffusion Model (TDM) for text recog nition which can guide IDM to generate text images with correct structures. We f urther propose a Mixture of Multi-modality module (MoM) to make these two diffus ion models cooperate with each other in all the diffusion steps. Extensive exper iments on synthetic and real-world datasets demonstrate that our Diffusion-based Blind Text Image Super-Resolution (DiffTSR) can restore text images with more a ccurate text structures as well as more realistic appearances simultaneously. Co de is available at https://github.com/YuzheZhang-1999/DiffTSR.
****
Consistent Prompting for Rehearsal-Free Continual Learning
Zhanxin Gao, Jun Cen, Xiaobin Chang; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2024, pp. 28463-28473
Continual learning empowers models to adapt autonomously to the ever-changing en vironment or data streams without forgetting old knowledge. Prompt-based approac hes are built on frozen pre-trained models to learn the task-specific prompts an d classifiers efficiently. Existing prompt based methods are inconsistent betwee n training and testing limiting their effectiveness. Two types of inconsistency are revealed. Test predictions are made from all classifiers while training only focuses on the current task classifier without holistic alignment leading to Cl assifier inconsistency. Prompt inconsistency indicates that the prompt selected during testing may not correspond to the one associated with this task during tr aining. In this paper we propose a novel prompt-based method Consistent Promptin g (CPrompt) for more aligned training and testing. Specifically all existing cla ssifiers are exposed to prompt training resulting in classifier consistency lear ning. In addition prompt consistency learning is proposed to enhance prediction robustness and boost prompt selection accuracy. Our Consistent Prompting surpass es its prompt-based counterparts and achieves state-of-the-art performance on mu ltiple continual learning benchmarks. Detailed analysis shows that improvements come from more consistent training and testing.
****
UniPAD: A Universal Pre-training Paradigm for Autonomous Driving
Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tan g, Hengshuang Zhao, Qibo Qiu, Binbin Lin, Xiaofei He, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 24, pp. 15238-15250
In the context of autonomous driving the significance of effective feature learn ing is widely acknowledged. While conventional 3D self-supervised pre-training m ethods have shown widespread success most methods follow the ideas originally de signed for 2D images. In this paper we present UniPAD a novel self-supervised le arning paradigm applying 3D volumetric differentiable rendering. UniPAD implicit ly encodes 3D space facilitating the reconstruction of continuous 3D shape struc tures and the intricate appearance characteristics of their 2D projections. The flexibility of our method enables seamless integration into both 2D and 3D frame works enabling a more holistic comprehension of the scenes. We manifest the feas ibility and effectiveness of UniPAD by conducting extensive experiments on vario us 3D perception tasks. Our method significantly improves lidar- camera- and lid ar-camera-based baseline by 9.1 7.7 and 6.9 NDS respectively. Notably our pre-tr aining pipeline achieves 73.2 NDS for 3D object detection and 79.4 mIoU for 3D s emantic segmentation on the nuScenes validation set achieving state-of-the-art r esults in comparison with previous methods.
****
SeD: Semantic-Aware Discriminator for Image Super-Resolution
Bingchen Li, Xin Li, Hanxin Zhu, Yeying Jin, Ruoyu Feng, Zhizheng Zhang, Zhibo C hen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2024, pp. 25784-25795
Generative Adversarial Networks (GANs) have been widely used to recover vivid te xtures in image super-resolution (SR) tasks. In particular one discriminator is utilized to enable the SR network to learn the distribution of real-world high-q uality images in an adversarial training manner. However the distribution learni ng is overly coarse-grained which is susceptible to virtual textures and causes

counter-intuitive generation results. To mitigate this we propose the simple and effective Semantic-aware Discriminator (denoted as SeD) which encourages the SR network to learn the fine-grained distributions by introducing the semantics of images as a condition. Concretely we aim to excavate the semantics of images from a well-trained semantic extractor. Under different semantics the discriminator is able to distinguish the real-fake images individually and adaptively which guides the SR network to learn the more fine-grained semantic-aware textures. To obtain accurate and abundant semantics we take full advantage of recently popular pretrained vision models (PVMs) with extensive datasets and then incorporate its semantic features into the discriminator through a well-designed spatial cross-attention module. In this way our proposed semantic-aware discriminator empowered the SR network to produce more photo-realistic and pleasing images. Extensive experiments on two typical tasks i.e. SR and Real SR have demonstrated the effectiveness of our proposed methods. The code will be available at https://github.com/lbc12345/SeD.
****
SocialCounterfactuals: Probing and Mitigating Intersectional Social Biases in Vision-Language Models with Counterfactual Examples
Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, Vasudev Lal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11975-11985
While vision-language models (VLMs) have achieved remarkable performance improvements recently there is growing evidence that these models also posses harmful biases with respect to social attributes such as gender and race. Prior studies have primarily focused on probing such bias attributes individually while ignoring biases associated with intersections between social attributes. This could be due to the difficulty of collecting an exhaustive set of image-text pairs for various combinations of social attributes. To address this challenge we employ text-to-image diffusion models to produce counterfactual examples for probing intersectional social biases at scale. Our approach utilizes Stable Diffusion with cross attention control to produce sets of counterfactual image-text pairs that are highly similar in their depiction of a subject (e.g. a given occupation) while differing only in their depiction of intersectional social attributes (e.g. race & gender). Through our over-generate-then-filter methodology we produce SocialCounterfactuals a high-quality dataset containing 171k image-text pairs for probing intersectional biases related to gender race and physical characteristics. We conduct extensive experiments to demonstrate the usefulness of our generated dataset for probing and mitigating intersectional social biases in state-of-the-art VLMs.
****
SVDTree: Semantic Voxel Diffusion for Single Image Tree Reconstruction
Yuan Li, Zhihao Liu, Bedrich Benes, Xiaopeng Zhang, Jianwei Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4692-4702
Efficiently representing and reconstructing the 3D geometry of biological trees remains a challenging problem in computer vision and graphics. We propose a novel approach for generating realistic tree models from single-view photographs. We cast the 3D information inference problem to a semantic voxel diffusion process which converts an input image of a tree to a novel Semantic Voxel Structure (SVS) in 3D space. The SVS encodes the geometric appearance and semantic structural information (e.g. classifying trunks branches and leaves) which retains the intricate internal tree features. Tailored to the SVS we present SVDTree a new hybrid tree modeling approach by combining structure-oriented branch reconstruction and self-organization-based foliage reconstruction. We validate SVDTree by using images from both synthetic and real trees. The comparison results show that our approach can better preserve tree details and achieve more realistic and accurate reconstruction results than previous methods.
****
Rethinking FID: Towards a Better Evaluation Metric for Image Generation
Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakr

abarti, Sanjiv Kumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9307-9315
As with many machine learning problems the progress of image generation methods hinges on good evaluation metrics. One of the most popular is the Frechet Inception Distance (FID). FID estimates the distance between a distribution of Inception-v3 features of real images and those of images generated by the algorithm. We highlight important drawbacks of FID: Inception's poor representation of the rich and varied content generated by modern text-to-image models incorrect normality assumptions and poor sample complexity. We call for a reevaluation of FID's use as the primary quality metric for generated images. We empirically demonstrate that FID contradicts human raters it does not reflect gradual improvement of iterative text-to-image models it does not capture distortion levels and that it produces inconsistent results when varying the sample size. We also propose an alternative new metric CMMD based on richer CLIP embeddings and the maximum mean discrepancy distance with the Gaussian RBF kernel. It is an unbiased estimator that does not make any assumptions on the probability distribution of the embeddings and is sample efficient. Through extensive experiments and analysis we demonstrate that FID-based evaluations of text-to-image models may be unreliable and that CMMD offers a more robust and reliable assessment of image quality.
****
Efficient Privacy-Preserving Visual Localization Using 3D Ray Clouds
Heejoon Moon, Chunghwan Lee, Je Hyeong Hong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9773-9783
The recent success in revealing scene details from sparse 3D point clouds obtained via structure-from-motion has raised significant privacy concerns in visual localization. One prominent approach for mitigating this issue is to lift 3D points to 3D lines thereby reducing the effectiveness of the scene inversion attacks but this comes at the cost of increased algorithmic complexity for camera localization due to weaker geometric constraints induced by line clouds. To overcome this limitation we propose a new lifting approach called "ray cloud" whereby each lifted 3D line intersects at one of two predefined locations depicting omnidirectional rays from two cameras. This yields two benefits i) camera localization can now be cast as relative pose estimation between the query image and the calibrated rig of two perspective cameras which can be efficiently solved using a variant of the 5-point algorithm and ii) the ray cloud introduces erroneous estimations for the density-based inversion attack degrading the quality of scene recovery. Moreover we explore possible modifications of the inversion attack to better recover scenes from the ray clouds and propose a ray sampling technique to reduce the effectiveness of the modified attack. Experimental results on two public datasets show real-time localization speed as well as enhanced privacy-preserving capability over the state-of-the-art without overly sacrificing the localization accuracy.
****
SuperPrimitive: Scene Reconstruction at a Primitive Level
Kirill Mazur, Gwangbin Bae, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4979-4989
Joint camera pose and dense geometry estimation from a set of images or a monocular video remains a challenging problem due to its computational complexity and inherent visual ambiguities. Most dense incremental reconstruction systems operate directly on image pixels and solve for their 3D positions using multi-view geometry cues. Such pixel-level approaches suffer from ambiguities or violations of multi-view consistency (e.g. caused by textureless or specular surfaces). We address this issue with a new image representation which we call a SuperPrimitive. SuperPrimitives are obtained by splitting images into semantically correlated local regions and enhancing them with estimated surface normal directions both of which are predicted by state-of-the-art single image neural networks. This provides a local geometry estimate per SuperPrimitive while their relative positions are adjusted based on multi-view observations. We demonstrate the versatility of our new representation by addressing three 3D reconstruction tasks: depth completion few-view structure from motion and monocular dense visual odometry. Proj

ect page: https://makezur.github.io/SuperPrimitive/
****
ReCoRe: Regularized Contrastive Representation Learning of World Model
Rudra P.K. Poudel, Harit Pandya, Stephan Liwicki, Roberto Cipolla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22904-22913
While recent model-free Reinforcement Learning (RL) methods have demonstrated human-level effectiveness in gaming environments their success in everyday tasks like visual navigation has been limited particularly under significant appearance variations. This limitation arises from (i) poor sample efficiency and (ii) over-fitting to training scenarios. To address these challenges we present a world model that learns invariant features using (i) contrastive unsupervised learning and (ii) an intervention-invariant regularizer. Learning an explicit representation of the world dynamics i.e. a world model improves sample efficiency while contrastive learning implicitly enforces learning of invariant features which improves generalization. However the naive integration of contrastive loss to world models is not good enough as world-model-based RL methods independently optimize representation learning and agent policy. To overcome this issue we propose an intervention-invariant regularizer in the form of an auxiliary task such as depth prediction image denoising image segmentation etc. that explicitly enforces invariance to style interventions. Our method outperforms current state-of-the-art model-based and model-free RL methods and significantly improves on out-of-distribution point navigation tasks evaluated on the iGibson benchmark. With only visual observations we further demonstrate that our approach outperforms recent language-guided foundation models for point navigation which is essential for deployment on robots with limited computation capabilities. Finally we demonstrate that our proposed model excels at the sim-to-real transfer of its perception module on the Gibson benchmark.
****
TFMQ-DM: Temporal Feature Maintenance Quantization for Diffusion Models
Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, Xianglong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7362-7371
The Diffusion model a prevalent framework for image generation encounters significant challenges in terms of broad applicability due to its extended inference times and substantial memory requirements. Efficient Post-training Quantization (PTQ) is pivotal for addressing these issues in traditional models. Different from traditional models diffusion models heavily depend on the time-step t to achieve satisfactory multi-round denoising. Usually t from the finite set \ 1 \ldots T\ is encoded to a temporal feature by a few modules totally irrespective of the sampling data. However existing PTQ methods do not optimize these modules separately. They adopt inappropriate reconstruction targets and complex calibration methods resulting in a severe disturbance of the temporal feature and denoising trajectory as well as a low compression efficiency. To solve these we propose a Temporal Feature Maintenance Quantization (TFMQ) framework building upon a Temporal Information Block which is just related to the time-step t and unrelated to the sampling data. Powered by the pioneering block design we devise temporal information aware reconstruction (TIAR) and finite set calibration (FSC) to align the full-precision temporal features in a limited time. Equipped with the framework we can maintain the most temporal information and ensure the end-to-end generation quality. Extensive experiments on various datasets and diffusion models prove our state-of-the-art results. Remarkably our quantization approach for the first time achieves model performance nearly on par with the full-precision model under 4-bit weight quantization. Additionally our method incurs almost no extra computational cost and accelerates quantization time by 2.0 xon LSUN-Bedrooms 256 x256 compared to previous works. Our code is publicly available at \href https://github.com/ModelTC/TFMQ-DM https://github.com/ModelTC/TFMQ-DM .
****
CNC-Net: Self-Supervised Learning for CNC Machining Operations
Mohsen Yavartanoo, Sangmin Hong, Reyhaneh Neshatavar, Kyoung Mu Lee; Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9816-9825
CNC manufacturing is a process that employs computer numerical control (CNC) machines to govern the movements of various industrial tools and machinery encompassing equipment ranging from grinders and lathes to mills and CNC routers. However the reliance on manual CNC programming has become a bottleneck and the requirement for expert knowledge can result in significant costs. Therefore we introduce a pioneering approach named CNC-Net representing the use of deep neural networks (DNNs) to simulate CNC machines and grasp intricate operations when supplied with raw materials. CNC-Net constitutes a self-supervised framework that exclusively takes an input 3D model and subsequently generates the essential operation parameters required by the CNC machine to construct the object. Our method has the potential to transformative automation in manufacturing by offering a cost-effective alternative to the high costs of manual CNC programming while maintaining exceptional precision in 3D object production. Our experiments underscore the effectiveness of our CNC-Net in constructing the desired 3D objects through the utilization of CNC operations. Notably it excels in preserving finer local details exhibiting a marked enhancement in precision compared to the state-of-the-art 3D CAD reconstruction approaches. The codes are available at https://github.com/myavartanoo/CNC-Net_PyTorch.
****
JRDB-PanoTrack: An Open-world Panoptic Segmentation and Tracking Robotic Dataset in Crowded Human Environments
Duy Tho Le, Chenhui Gou, Stavya Datta, Hengcan Shi, Ian Reid, Jianfei Cai, Hamid Rezatofighi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22325-22334
Autonomous robot systems have attracted increasing research attention in recent years where environment understanding is a crucial step for robot navigation human-robot interaction and decision. Real-world robot systems usually collect visual data from multiple sensors and are required to recognize numerous objects and their movements in complex human-crowded settings. Traditional benchmarks with their reliance on single sensors and limited object classes and scenarios fail to provide the comprehensive environmental understanding robots need for accurate navigation interaction and decision-making. As an extension of JRDB dataset we unveil JRDB-PanoTrack a novel open-world panoptic segmentation and tracking benchmark towards more comprehensive environmental perception. JRDB-PanoTrack includes (1) various data involving indoor and outdoor crowded scenes as well as comprehensive 2D and 3D synchronized data modalities; (2) high-quality 2D spatial panoptic segmentation and temporal tracking annotations with additional 3D label projections for further spatial understanding; (3) diverse object classes for closed- and open-world recognition benchmarks with OSPA-based metrics for evaluation. Extensive evaluation of leading methods shows significant challenges posed by our dataset.
****
CONFORM: Contrast is All You Need for High-Fidelity Text-to-Image Diffusion Models
Tuna Han Salih Meral, Enis Simsar, Federico Tombari, Pinar Yanardag; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9005-9014
Images produced by text-to-image diffusion models might not always faithfully represent the semantic intent of the provided text prompt where the model might overlook or entirely fail to produce certain objects. While recent studies propose various solutions they often require customly tailored functions for each of these problems leading to sub-optimal results especially for complex prompts. Our work introduces a novel perspective by tackling this challenge in a contrastive context. Our approach intuitively promotes the segregation of objects in attention maps while also maintaining that pairs of related attributes are kept close to each other. We conducted extensive experiments across a wide variety of scenarios each involving unique combinations of objects attributes and scenes. These experiments effectively showcase the versatility efficiency and flexibility of ou

r method in working with both latent and pixel-based diffusion models including Stable Diffusion and Imagen. Moreover we publicly share our source code to facilitate further research.
****
Self-Supervised Facial Representation Learning with Facial Region Awareness
Zheng Gao, Ioannis Patras; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2081-2092
Self-supervised pre-training has been proved to be effective in learning transferable representations that benefit various visual tasks. This paper asks this question: can self-supervised pre-training learn general facial representations for various facial analysis tasks? Recent efforts toward this goal are limited to treating each face image as a whole i.e. learning consistent facial representations at the image-level which overlooks the consistency of local facial representations (i.e. facial regions like eyes nose etc). In this work we make a first attempt to propose a novel self-supervised facial representation learning framework to learn consistent global and local facial representations Facial Region Awareness (FRA). Specifically we explicitly enforce the consistency of facial regions by matching the local facial representations across views which are extracted with learned heatmaps highlighting the facial regions. Inspired by the mask prediction in supervised semantic segmentation we obtain the heatmaps via cosine similarity between the per-pixel projection of feature maps and facial mask embeddings computed from learnable positional embeddings which leverage the attention mechanism to globally look up the facial image for facial regions. To learn such heatmaps we formulate the learning of facial mask embeddings as a deep clustering problem by assigning the pixel features from the feature maps to them. The transfer learning results on facial classification and regression tasks show that our FRA outperforms previous pre-trained models and more importantly using ResNet as the unified backbone for various tasks our FRA achieves comparable or even better performance compared with SOTA methods in facial analysis tasks.
****
GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models
Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, Xinggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6796-6807
In recent times the generation of 3D assets from text prompts has shown impressive results. Both 2D and 3D diffusion models can help generate decent 3D objects based on prompts. 3D diffusion models have good 3D consistency but their quality and generalization are limited as trainable 3D data is expensive and hard to obtain. 2D diffusion models enjoy strong abilities of generalization and fine generation but 3D consistency is hard to guarantee. This paper attempts to bridge the power from the two types of diffusion models via the recent explicit and efficient 3D Gaussian splatting representation. A fast 3D object generation framework named as GaussianDreamer is proposed where the 3D diffusion model provides priors for initialization and the 2D diffusion model enriches the geometry and appearance. Operations of noisy point growing and color perturbation are introduced to enhance the initialized Gaussians. Our GaussianDreamer can generate a high-quality 3D instance or 3D avatar within 15 minutes on one GPU much faster than previous methods while the generated instances can be directly rendered in real time. Demos and code are available at https://taoranyi.com/gaussiandreamer/.
****
Open-Vocabulary Attention Maps with Token Optimization for Semantic Segmentation in Diffusion Models
Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C. SanMiguel, José M. Martínez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9242-9252
Diffusion models represent a new paradigm in text-to-image generation. Beyond generating high-quality images from text prompts models such as Stable Diffusion have been successfully extended to the joint generation of semantic segmentation pseudo-masks. However current extensions primarily rely on extracting attentions

linked to prompt words used for image synthesis. This approach limits the generation of segmentation masks derived from word tokens not contained in the text prompt. In this work we introduce Open-Vocabulary Attention Maps (OVAM)--a training-free method for text-to-image diffusion models that enables the generation of attention maps for any word. In addition we propose a lightweight optimization process based on OVAM for finding tokens that generate accurate attention maps for an object class with a single annotation. We evaluate these tokens within existing state-of-the-art Stable Diffusion extensions. The best-performing model improves its mIoU from 52.1 to 86.6 for the synthetic images' pseudo-masks demonstrating that our optimized tokens are an efficient way to improve the performance of existing methods without architectural changes or retraining.
****

OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13418-13427

Hallucination posed as a pervasive challenge of multi-modal large language models (MLLMs) has significantly impeded their real-world usage that demands precise judgment. Existing methods mitigate this issue with either training with specific designed data or inferencing with external knowledge from other sources incurring inevitable additional costs. In this paper we present OPERA a novel MLLM decoding method grounded in an Over-trust Penalty and a Retrospection-Allocation strategy serving as a nearly free lunch to alleviate the hallucination issue without additional data knowledge or training. Our approach begins with an interesting observation that most hallucinations are closely tied to the knowledge aggregation patterns manifested in the self-attention matrix i.e. MLLMs tend to generate new tokens by focusing on a few summary tokens but not all the previous tokens. Such partial over-trust inclination results in the neglecting of image tokens and describes the image content with hallucination. Based on the observation OPERA introduces a penalty term on the model logits during the beam-search decoding to mitigate the over-trust issue along with a rollback strategy that retrospects the presence of summary tokens in the previously generated tokens and re-allocate the token selection if necessary. With extensive experiments OPERA shows significant hallucination-mitigating performance on different MLLMs and metrics proving its effectiveness and generality. Our code is available at: https://github.com/shikiw/OPERA.
****

Volumetric Environment Representation for Vision-Language Navigation

Rui Liu, Wenguan Wang, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16317-16328

Vision-language navigation (VLN) requires an agent to navigate through an 3D environment based on visual observations and natural language instructions. It is clear that the pivotal factor for successful navigation lies in the comprehensive scene understanding. Previous VLN agents employ monocular frameworks to extract 2D features of perspective views directly. Though straightforward they struggle for capturing 3D geometry and semantics leading to a partial and incomplete environment representation. To achieve a comprehensive 3D representation with fine-grained details we introduce a Volumetric Environment Representation (VER) which voxelizes the physical world into structured 3D cells. For each cell VER aggregates multi-view 2D features into such a unified 3D space via 2D-3D sampling. Through coarse-to-fine feature extraction and multi-task learning for VER our agent predicts 3D occupancy 3D room layout and 3D bounding boxes jointly. Based on online collected VERs our agent performs volume state estimation and builds episodic memory for predicting the next step. Experimental results show our environment representations from multi-task learning lead to evident performance gains on VLN. Our model achieves state-of-the-art performance across VLN benchmarks (R2R REVERIE and R4R).
****

DreamComposer: Controllable 3D Object Generation via Multi-View Conditions

Yunhan Yang, Yukun Huang, Xiaoyang Wu, Yuan-Chen Guo, Song-Hai Zhang, Hengshuang Zhao, Tong He, Xihui Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8111-8120
Utilizing pre-trained 2D large-scale generative models recent works are capable of generating high-quality novel views from a single in-the-wild image. However due to the lack of information from multiple views these works encounter difficulties in generating controllable novel views. In this paper we present DreamComposer a flexible and scalable framework that can enhance existing view-aware diffusion models by injecting multi-view conditions. Specifically DreamComposer first uses a view-aware 3D lifting module to obtain 3D representations of an object from multiple views. Then it renders the latent features of the target view from 3D representations with the multi-view feature fusion module. Finally the target view features extracted from multi-view inputs are injected into a pre-trained diffusion model. Experiments show that DreamComposer is compatible with state-of-the-art diffusion models for zero-shot novel view synthesis further enhancing them to generate high-fidelity novel view images with multi-view conditions ready for controllable 3D object reconstruction and various other applications.
****
Self-Calibrating Vicinal Risk Minimisation for Model Calibration
Jiawei Liu, Changkun Ye, Ruikai Cui, Nick Barnes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3335-3345
Model calibration measuring the alignment between the prediction accuracy and model confidence is an important metric reflecting model trustworthiness. Existing dense binary classification methods without proper regularisation of model confidence are prone to being over-confident. To calibrate Deep Neural Networks (DNNs) we propose a Self-Calibrating Vicinal Risk Minimisation (SCVRM) that explores the vicinity space of labeled data where vicinal images that are farther away from labeled images adopt the groundtruth label with decreasing label confidence. We prove that in the logistic regression problem SCVRM can be seen as a Vicinal Risk Minimisation plus a regularisation term that penalises the over-confident predictions. In practical implementation SCVRM is approximated using Monte Carlo sampling that samples additional augmented training images and labels from the vicinal distributions. Experimental results demonstrate that SCVRM can significantly enhance model calibration for different dense classification tasks on both in-distribution and out-of-distribution data. Code is available at https://github.com/Carlisle-Liu/SCVRM.
****
NeRFDeformer: NeRF Transformation from a Single View via 3D Scene Flows
Zhenggang Tang, Zhongzheng Ren, Xiaoming Zhao, Bowen Wen, Jonathan Tremblay, Stan Birchfield, Alexander Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10293-10303
We present a method for automatically modifying a NeRF representation based on a single observation of a non-rigid transformed version of the original scene. Our method defines the transformation as a 3D flowspecifically as a weighted linear blending of rigid transformations of 3D anchor points that are defined on the surface of the scene. In order to identify anchor points we introduce a novel correspondence algorithm that first matches RGB-based pairs then leverages multi-view information and 3D reprojection to robustly filter false positives in two steps. We also introduce a new dataset for exploring the problem of modifying a NeRF scene through a single observation. Our dataset contains 113 scenes leveraging 47 3D assets.We show that our proposed method outperforms NeRF editing methods as well as diffusion-based methods and we also explore different methods for filtering correspondences.
****
LPSNet: End-to-End Human Pose and Shape Estimation with Lensless Imaging
Haoyang Ge, Qiao Feng, Hailong Jia, Xiongzheng Li, Xiangjun Yin, You Zhou, Jingyu Yang, Kun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1471-1480
Human pose and shape (HPS) estimation with lensless imaging is not only beneficial to privacy protection but also can be used in covert surveillance scenarios d

ue to the small size and simple structure of this device. However this task pres
ents significant challenges due to the inherent ambiguity of the captured measur
ements and lacks effective methods for directly estimating human pose and shape
from lensless data. In this paper we propose the first end-to-end framework to r
ecover 3D human poses and shapes from lensless measurements to our knowledge. We
 specifically design a multi-scale lensless feature decoder to decode the lensle
ss measurements through the optically encoded mask for efficient feature extract
ion. We also propose a double-head auxiliary supervision mechanism to improve th
e estimation accuracy of human limb ends. Besides we establish a lensless imagin
g system and verify the effectiveness of our method on various datasets acquired
 by our lensless imaging system. The code and dataset are available at https://c
ic.tju.edu.cn/faculty/likun/projects/LPSNet.
****
Embracing Unimodal Aleatoric Uncertainty for Robust Multimodal Fusion
Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, Heng Tao Shen; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 26876-26885
As a fundamental problem in multimodal learning multimodal fusion aims to compen
sate for the inherent limitations of a single modality. One challenge of multimo
dal fusion is that the unimodal data in their unique embedding space mostly cont
ains potential noise which leads to corrupted cross-modal interactions. However
in this paper we show that the potential noise in unimodal data could be well qu
antified and further employed to enhance more stable unimodal embeddings via con
trastive learning. Specifically we propose a novel generic and robust multimodal
 fusion strategy termed Embracing Aleatoric Uncertainty (EAU) which is simple an
d can be applied to kinds of modalities. It consists of two key steps: (1) the S
table Unimodal Feature Augmentation (SUFA) that learns a stable unimodal represe
ntation by incorporating the aleatoric uncertainty into self-supervised contrast
ive learning. (2) Robust Multimodal Feature Integration (RMFI) leveraging an inf
ormation-theoretic strategy to learn a robust compact joint representation. We e
valuate our proposed EAU method on five multimodal datasets where the video RGB
image text audio and depth image are involved. Extensive experiments demonstrate
 the EAU method is more noise-resistant than existing multimodal fusion strategi
es and establishes new state-of-the-art on several benchmarks.
****
Unifying Correspondence Pose and NeRF for Generalized Pose-Free Novel View Synth
esis
Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jiaolong Yang, Seungryong Kim, Chong
Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2024, pp. 20196-20206
This work delves into the task of pose-free novel view synthesis from stereo pai
rs a challenging and pioneering task in 3D vision. Our innovative framework unli
ke any before seamlessly integrates 2D correspondence matching camera pose estim
ation and NeRF rendering fostering a synergistic enhancement of these tasks. We
achieve this through designing an architecture that utilizes a shared representa
tion which serves as a foundation for enhanced 3D geometry understanding. Capita
lizing on the inherent interplay between the tasks our unified framework is trai
ned end-to-end with the proposed training strategy to improve overall model accu
racy. Through extensive evaluations across diverse indoor and outdoor scenes fro
m two real-world datasets we demonstrate that our approach achieves substantial
improvement over previous methodologies especially in scenarios characterized by
 extreme viewpoint changes and the absence of accurate camera poses.
****
Draw Step by Step: Reconstructing CAD Construction Sequences from Point Clouds v
ia Multimodal Diffusion.
Weijian Ma, Shuaiqi Chen, Yunzhong Lou, Xueyang Li, Xiangdong Zhou; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 27154-27163
Reconstructing CAD construction sequences from raw 3D geometry serves as an inte
rface between real-world objects and digital designs. In this paper we propose C

AD-Diffuser a multimodal diffusion scheme aiming at integrating top-down design paradigm into generative reconstruction. In particular we unify CAD point clouds and CAD construction sequences at the token level guiding our proposed multimodal diffusion strategy to understand and link between the geometry and the design intent concentrated in construction sequences. Leveraging the strong decoding abilities of language models the forward process is modeled as a random walk between the original token and the [MASK] token while the reverse process naturally fits the masked token modeling scheme. A volume-based noise schedule is designed to encourage outline-first generation decomposing the top-down design methodology into a machine-understandable procedure. For tokenizing CAD data of multiple modalities we introduce a tokenizer with a self-supervised face segmentation task to compress local and global geometric information for CAD point clouds and the CAD construction sequence is transformed into a primitive token string. Experimental results show that our CAD-Diffuser can perceive geometric details and the results are more likely to be reused by human designers.
****
DiffusionTrack: Point Set Diffusion Model for Visual Object Tracking
Fei Xie, Zhongdao Wang, Chao Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19113-19124
Existing Siamese or transformer trackers commonly pose visual object tracking as a one-shot detection problem i.e. locating the target object in a single forward evaluation scheme. Despite the demonstrated success these trackers may easily drift towards distractors with similar appearance due to the single forward evaluation scheme lacking self-correction. To address this issue we cast visual tracking as a point set based denoising diffusion process and propose a novel generative learning based tracker dubbed DiffusionTrack. Our DiffusionTrack possesses two appealing properties: 1) It follows a novel noise-to-target tracking paradigm that leverages multiple denoising diffusion steps to localize the target in a dynamic searching manner per frame. 2) It models the diffusion process using a point set representation which can better handle appearance variations for more precise localization. One side benefit is that DiffusionTrack greatly simplifies the post-processing e.g. removing window penalty scheme. Without bells and whistles our DiffusionTrack achieves leading performance over the state-of-the-art trackers and runs in real-time. The code is in https://github.com/VISION-SJTU/DiffusionTrack.
****
Towards a Simultaneous and Granular Identity-Expression Control in Personalized Face Generation
Renshuai Liu, Bowen Ma, Wei Zhang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Xuan Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2114-2123
In human-centric content generation the pre-trained text-to-image models struggle to produce user-wanted portrait images which retain the identity of individuals while exhibiting diverse expressions. This paper introduces our efforts towards personalized face generation. To this end we propose a novel multi-modal face generation framework capable of simultaneous identity-expression control and more fine-grained expression synthesis. Our expression control is so sophisticated that it can be specialized by the fine-grained emotional vocabulary. We devise a novel diffusion model that can undertake the task of simultaneously face swapping and reenactment. Due to the entanglement of identity and expression separately and precisely controlling them within one framework is a nontrivial task thus has not been explored yet. To overcome this we propose several innovative designs in the conditional diffusion model including balancing identity and expression encoder improved midpoint sampling and explicitly background conditioning. Extensive experiments have demonstrated the controllability and scalability of the proposed framework in comparison with state-of-the-art text-to-image face swapping and face reenactment methods.
****
PEEKABOO: Interactive Video Generation via Masked-Diffusion
Yash Jain, Anshul Nasery, Vibhav Vineet, Harkirat Behl; Proceedings of the IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8079-8088
Modern video generation models like Sora have achieved remarkable success in producing high-quality videos. However a significant limitation is their inability to offer interactive control to users a feature that promises to open up unprecedented applications and creativity. In this work we introduce the first solution to equip diffusion-based video generation models with spatio-temporal control. We present Peekaboo a novel masked attention module which seamlessly integrates with current video generation models offering control without the need for additional training or inference overhead. To facilitate future research we also introduce a comprehensive benchmark for interactive video generation. This benchmark offers a standardized framework for the community to assess the efficacy of emerging interactive video generation models. Our extensive qualitative and quantitative assessments reveal that Peekaboo achieves up to a 3.8x improvement in mIoU over baseline models all while maintaining the same latency. Code and benchmark are available on the webpage.
****
Scaling Diffusion Models to Real-World 3D LiDAR Scene Completion
Lucas Nunes, Rodrigo Marcuzzi, Benedikt Mersch, Jens Behley, Cyrill Stachniss; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14770-14780
Computer vision techniques play a central role in the perception stack of autonomous vehicles. Such methods are employed to perceive the vehicle surroundings given sensor data. 3D LiDAR sensors are commonly used to collect sparse 3D point clouds from the scene. However compared to human perception such systems struggle to deduce the unseen parts of the scene given those sparse point clouds. In this matter the scene completion task aims at predicting the gaps in the LiDAR measurements to achieve a more complete scene representation. Given the promising results of recent diffusion models as generative models for images we propose extending them to achieve scene completion from a single 3D LiDAR scan. Previous works used diffusion models over range images extracted from LiDAR data directly applying image-based diffusion methods. Distinctly we propose to directly operate on the points reformulating the noising and denoising diffusion process such that it can efficiently work at scene scale. Together with our approach we propose a regularization loss to stabilize the noise predicted during the denoising process. Our experimental evaluation shows that our method can complete the scene given a single LiDAR scan as input producing a scene with more details compared to state-of-the-art scene completion methods. We believe that our proposed diffusion process formulation can support further research in diffusion models applied to scene-scale point cloud data.
****
Discriminative Pattern Calibration Mechanism for Source-Free Domain Adaptation
Haifeng Xia, Siyu Xia, Zhengming Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23648-23658
Source-free domain adaptation (SFDA) assumes that model adaptation only accesses the well-learned source model and unlabeled target instances for knowledge transfer. However cross-domain distribution shift easily triggers invalid discriminative semantics from source model on recognizing the target samples. Hence understanding the specific content of discriminative pattern and adjusting their representation in target domain become the important key to overcome SFDA. To achieve such a vision this paper proposes a novel explanation paradigm "Discriminative Pattern Calibration (DPC)" mechanism on solving SFDA issue. Concretely DPC first utilizes learning network to infer the discriminative regions on the target images and specifically emphasizes them in feature space to enhance their representation. Moreover DPC relies on the attention-reversed mixup mechanism to augment more samples and improve the robustness of the classifier. Considerable experimental results and studies suggest that the effectiveness of our DPC in enhancing the performance of existing SFDA baselines.
****
Deep Generative Model based Rate-Distortion for Image Downscaling Assessment

Yuanbang Liang, Bhavesh Garg, Paul Rosin, Yipeng Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19363-19372
In this paper we propose Image Downscaling Assessment by Rate-Distortion (IDA-RD) a novel measure to quantitatively evaluate image downscaling algorithms. In contrast to image-based methods that measure the quality of downscaled images ours is process-based that draws ideas from rate-distortion theory to measure the distortion incurred during downscaling. Our main idea is that downscaling and super-resolution (SR) can be viewed as the encoding and decoding processes in the rate-distortion model respectively and that a downscaling algorithm that preserves more details in the resulting low-resolution (LR) images should lead to less distorted high-resolution (HR) images in SR. In other words the distortion should increase as the downscaling algorithm deteriorates. However it is non-trivial to measure this distortion as it requires the SR algorithm to be blind and stochastic. Our key insight is that such requirements can be met by recent SR algorithms based on deep generative models that can find all matching HR images for a given LR image on their learned image manifolds. Extensive experimental results show the effectiveness of our IDA-RD measure.
****
Physical Backdoor: Towards Temperature-based Backdoor Attacks in the Physical World
Wen Yin, Jian Lou, Pan Zhou, Yulai Xie, Dan Feng, Yuhua Sun, Tailai Zhang, Lichao Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12733-12743
Backdoor attacks have been well-studied in visible light object detection (VLOD) in recent years. However VLOD can not effectively work in dark and temperature-sensitive scenarios. Instead thermal infrared object detection (TIOD) is the most accessible and practical in such environments. In this paper our team is the first to investigate the security vulnerabilities associated with TIOD in the context of backdoor attacks spanning both the digital and physical realms. We introduce two novel types of backdoor attacks on TIOD each offering unique capabilities: Object-affecting Attack and Range-affecting Attack. We conduct a comprehensive analysis of key factors influencing trigger design which include temperature size material and concealment. These factors especially temperature significantly impact the efficacy of backdoor attacks on TIOD. A thorough understanding of these factors will serve as a foundation for designing physical triggers and temperature controlling experiments. Our study includes extensive experiments conducted in both digital and physical environments. In the digital realm we evaluate our approach using benchmark datasets for TIOD achieving an Attack Success Rate (ASR) of up to 98.21%. In the physical realm we test our approach in two real-world settings: a traffic intersection and a parking lot using a thermal infrared camera. Here we attain an ASR of up to 98.38%.
****
Make Me a BNN: A Simple Strategy for Estimating Bayesian Uncertainty from Pre-trained Models
Gianni Franchi, Olivier Laurent, Maxence Leguery, Andrei Bursuc, Andrea Pilzer, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12194-12204
Deep Neural Networks (DNNs) are powerful tools for various computer vision tasks yet they often struggle with reliable uncertainty quantification -a critical requirement for real-world applications. Bayesian Neural Networks (BNN) are equipped for uncertainty estimation but cannot scale to large DNNs where they are highly unstable to train. To address this challenge we introduce the Adaptable Bayesian Neural Network (ABNN) a simple and scalable strategy to seamlessly transform DNNs into BNNs in a post-hoc manner with minimal computational and training overheads. ABNN preserves the main predictive properties of DNNs while enhancing their uncertainty quantification abilities through simple BNN adaptation layers (attached to normalization layers) and a few fine-tuning steps on pre-trained models. We conduct extensive experiments across multiple datasets for image classification and semantic segmentation tasks and our results demonstrate that ABNN ach

ieves state-of-the-art performance without the computational budget typically associated with ensemble methods.
****

Language-only Training of Zero-shot Composed Image Retrieval
Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, Sangdoo Yun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13225-13234
Composed image retrieval (CIR) task takes a composed query of image and text aiming to search relative images for both conditions. Conventional CIR approaches need a training dataset composed of triplets of query image query text and target image which is very expensive to collect. Several recent works have worked on the zero-shot (ZS) CIR paradigm to tackle the issue without using pre-collected triplets. However the existing ZS-CIR methods show limited backbone scalability and generalizability due to the lack of diversity of the input texts during training. We propose a novel CIR framework only using language for its training. Our LinCIR (Language-only training for CIR) can be trained only with text datasets by a novel self-supervision named self-masking projection (SMP). We project the text latent embedding to the token embedding space and construct a new text by replacing the keyword tokens of the original text. Then we let the new and original texts have the same latent embedding vector. With this simple strategy LinCIR is surprisingly efficient and highly effective; LinCIR with CLIP ViT-G backbone is trained in 48 minutes and shows the best ZS-CIR performances on four different CIR benchmarks CIRCO GeneCIS FashionIQ and CIRR even outperforming supervised method on FashionIQ. Code is available at https://github.com/navervision/lincir
****

EFHQ: Multi-purpose ExtremePose-Face-HQ dataset
Trung Tuan Dao, Duc Hong Vu, Cuong Pham, Anh Tran; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22605-22615
The existing facial datasets while having plentiful images at near frontal views lack images with extreme head poses leading to the downgraded performance of deep learning models when dealing with profile or pitched faces. This work aims to address this gap by introducing a novel dataset named Extreme Pose Face High-Quality Dataset (EFHQ) which includes a maximum of 450k high-quality images of faces at extreme poses. To produce such a massive dataset we utilize a novel and meticulous dataset processing pipeline to curate two publicly available datasets VFHQ and CelebV-HQ which contain many high-resolution face videos captured in various settings. Our dataset can complement existing datasets on various facial-related tasks such as facial synthesis with 2D/3D-aware GAN diffusion-based text-to-image face generation and face reenactment. Specifically training with EFHQ helps models generalize well across diverse poses significantly improving performance in scenarios involving extreme views confirmed by extensive experiments. Additionally we utilize EFHQ to define a challenging cross-view face verification benchmark in which the performance of SOTA face recognition models drops 5-37% compared to frontal-to-frontal scenarios aiming to stimulate studies on face recognition under severe pose conditions in the wild.
****

Dynamic Cues-Assisted Transformer for Robust Point Cloud Registration
Hong Chen, Pei Yan, Sihe Xiang, Yihua Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21698-21707
Point Cloud Registration is a critical and challenging task in computer vision. Recent advancements have predominantly embraced a coarse-to-fine matching mechanism with the key to matching the superpoints located in patches with inter-frame consistent structures. However previous methods still face challenges with ambiguous matching because the interference information aggregated from irrelevant regions may disturb the capture of inter-frame consistency relations leading to wrong matches. To address this issue we propose Dynamic Cues-Assisted Transformer (DCATr). Firstly the interference from irrelevant regions is greatly reduced by constraining attention to certain cues i.e. regions with highly correlated structures of potential corresponding superpoints. Secondly cues-assisted attention

is designed to mine the inter-frame consistency relations while more attention is assigned to pairs with high consistent confidence in feature aggregation. Finally a dynamic updating fashion is proposed to facilitate mining richer consistency information further improving aggregated features' distinctiveness and relieving matching ambiguity. Extensive evaluations on indoor and outdoor standard benchmarks demonstrate that DCATr outperforms all state-of-the-art methods.
****
Patch2Self2: Self-supervised Denoising on Coresets via Matrix Sketching
Shreyas Fadnavis, Agniva Chowdhury, Joshua Batson, Petros Drineas, Eleftherios Garyfallidis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27641-27651
Diffusion MRI (dMRI) non-invasively maps brain white matter yet necessitates denoising due to low signal-to-noise ratios. Patch2Self (P2S) employing self-supervised techniques and regression on a Casorati matrix effectively denoises dMRI images and has become the new de-facto standard in this field. P2S however is resource intensive both in terms of running time and memory usage as it uses all voxels (n) from all-but-one held-in volumes (d-1) to learn a linear mapping Phi : \mathbb R ^ n x(d-1) \mapsto \mathbb R ^ n for denoising the held-out volume. The increasing size and dimensionality of higher resolution dMRI acquisitions can make P2S infeasible for large-scale analyses. This work exploits the redundancy imposed by P2S to alleviate its performance issues and inspect regions that influence the noise disproportionately. Specifically this study makes a three-fold contribution: (1) We present Patch2Self2 (P2S2) a method that uses matrix sketching to perform self-supervised denoising. By solving a sub-problem on a smaller sub-space so called coreset we show how P2S2 can yield a significant speedup in training time while using less memory. (2) We present a theoretical analysis of P2S2 focusing on determining the optimal sketch size through rank estimation a key step in achieving a balance between denoising accuracy and computational efficiency. (3) We show how the so-called statistical leverage scores can be used to interpret the denoising of dMRI data a process that was traditionally treated as a black-box. Experimental results on both simulated and real data affirm that P2S2 maintains denoising quality while significantly enhancing speed and memory efficiency achieved by training on a reduced data subset.
****
High-fidelity Person-centric Subject-to-Image Synthesis
Yibin Wang, Weizhong Zhang, Jianwei Zheng, Cheng Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7675-7684
Current subject-driven image generation methods encounter significant challenges in person-centric image generation. The reason is that they learn the semantic scene and person generation by fine-tuning a common pre-trained diffusion which involves an irreconcilable training imbalance. Precisely to generate realistic persons they need to sufficiently tune the pre-trained model which inevitably causes the model to forget the rich semantic scene prior and makes scene generation over-fit to the training data. Moreover even with sufficient fine-tuning these methods can still not generate high-fidelity persons since joint learning of the scene and person generation also lead to quality compromise. In this paper we propose Face-diffuser an effective collaborative generation pipeline to eliminate the above training imbalance and quality compromise. Specifically we first develop two specialized pre-trained diffusion models i.e. Text-driven Diffusion Model (TDM) and Subject-augmented Diffusion Model (SDM) for scene and person generation respectively. The sampling process is divided into three sequential stages i.e. semantic scene construction subject-scene fusion and subject enhancement. The first and last stages are performed by TDM and SDM respectively. The subject-scene fusion stage that is the collaboration achieved through a novel and highly effective mechanism Saliency-adaptive Noise Fusion (SNF). Specifically it is based on our key observation that there exists a robust link between classifier-free guidance responses and the saliency of generated images. In each time step SNF leverages the unique strengths of each model and allows for the spatial blending of predicted noises from both models automatically in a saliency-aware manner

all of which can be seamlessly integrated into the DDIM sampling process. Extensive experiments confirm the impressive effectiveness and robustness of the Face-diffuser in generating high-fidelity person images depicting multiple unseen persons with varying contexts. Code is available at https://github.com/CodeGoat24/Face-diffuser.
****
The Devil is in the Fine-Grained Details: Evaluating Open-Vocabulary Object Detectors for Fine-Grained Understanding
Lorenzo Bianchi, Fabio Carrara, Nicola Messina, Claudio Gennaro, Fabrizio Falchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22520-22529
Recent advancements in large vision-language models enabled visual object detection in open-vocabulary scenarios where object classes are defined in free-text formats during inference. In this paper we aim to probe the state-of-the-art methods for open-vocabulary object detection to determine to what extent they understand fine-grained properties of objects and their parts. To this end we introduce an evaluation protocol based on dynamic vocabulary generation to test whether models detect discern and assign the correct fine-grained description to objects in the presence of hard-negative classes. We contribute with a benchmark suite of increasing difficulty and probing different properties like color pattern and material. We further enhance our investigation by evaluating several state-of-the-art open-vocabulary object detectors using the proposed protocol and find that most existing solutions which shine in standard open-vocabulary benchmarks struggle to accurately capture and distinguish finer object details. We conclude the paper by highlighting the limitations of current methodologies and exploring promising research directions to overcome the discovered drawbacks. Data and code are available at https://lorebianchi98.github.io/FG-OVD .
****
Efficient and Effective Weakly-Supervised Action Segmentation via Action-Transition-Aware Boundary Alignment
Angchi Xu, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18253-18262
Weakly-supervised action segmentation is a task of learning to partition a long video into several action segments where training videos are only accompanied by transcripts (ordered list of actions). Most of existing methods need to infer pseudo segmentation for training by serial alignment between all frames and the transcript which is time-consuming and hard to be parallelized while training. In this work we aim to escape from this inefficient alignment with massive but redundant frames and instead to directly localize a few action transitions for pseudo segmentation generation where a transition refers to the change from an action segment to its next adjacent one in the transcript. As the true transitions are submerged in noisy boundaries due to intra-segment visual variation we propose a novel Action-Transition-Aware Boundary Alignment (ATBA) framework to efficiently and effectively filter out noisy boundaries and detect transitions. In addition to boost the semantic learning in the case that noise is inevitably present in the pseudo segmentation we also introduce video-level losses to utilize the trusted video-level supervision. Extensive experiments show the effectiveness of our approach on both performance and training speed.
****
Link-Context Learning for Multimodal LLMs
Yan Tai, Weichen Fan, Zhao Zhang, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27176-27185
The ability to learn from context with novel concepts and deliver appropriate responses are essential in human conversations. Despite current Multimodal Large Language Models (MLLMs) and Large Language Models (LLMs) being trained on mega-scale datasets recognizing unseen images or understanding novel concepts in a training-free manner remains a challenge. In-Context Learning (ICL) explores training-free few-shot learning where models are encouraged to "learn to learn" from limited tasks and generalize to unseen tasks. In this work we propose link-context learning (LCL) which emphasizes "reasoning from cause and effect" to augment th

e learning capabilities of MLLMs. LCL goes beyond traditional ICL by explicitly strengthening the causal relationship between the support set and the query set. By providing demonstrations with causal links LCL guides the model to discern not only the analogy but also the underlying causal associations between data points which empowers MLLMs to recognize unseen images and understand novel concepts more effectively. To facilitate the evaluation of this novel approach we introduce the ISEKAI dataset comprising exclusively of unseen generated image-label pairs designed for link-context learning. Extensive experiments show that our LCL-MLLM exhibits strong link-context learning capabilities to novel concepts over vanilla MLLMs.
****
Pixel-Aligned Language Model
Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13030-13039
Large language models have achieved great success in recent years so as their variants in vision. Existing vision-language models can describe images in natural languages answer visual-related questions or perform complex reasoning about the image. However it is yet unclear how localization tasks such as word grounding or referring localization can be performed using large language models. In this work we aim to develop a vision-language model that can take locations for example a set of points or boxes as either inputs or outputs. When taking locations as inputs the model performs location-conditioned captioning which generates captions for the indicated object or region. When generating locations as outputs our model regresses pixel coordinates for each output word generated by the language model and thus performs dense word grounding. Our model is pre-trained on the Localized Narrative dataset which contains pixel-word-aligned captioning from human attention. We show our model can be applied to various location-aware vision-language tasks including referring localization location-conditioned captioning and dense object captioning archiving state-of-the-art performance on RefCOCO and Visual Genome.
****
JeDi: Joint-Image Diffusion Models for Finetuning-Free Personalized Text-to-Image Generation
Yu Zeng, Vishal M. Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, Yogesh Balaji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6786-6795
Personalized text-to-image generation models enable users to create images that depict their individual possessions in diverse scenes finding applications in various domains. To achieve the personalization capability existing methods rely on finetuning a text-to-image foundation model on a user's custom dataset which can be non-trivial for general users resource-intensive and time-consuming. Despite attempts to develope finetuning-free methods their generation quality is much lower compared to their finetuning counterparts. In this paper we propose Joint-Image Diffusion (\jedi) an effective technique for learning a finetuning-free personalization model. Our key idea is to learn the joint distribution of multiple related text-image pairs that share a common subject. To facilitate learning we propose a scalable synthetic dataset generation technique. Once trained our model enables fast and easy personalization at test time by simply using reference images as input during the sampling process. Our approach does not require any expensive optimization process or additional modules and can faithfully preserve the identity represented by any number of reference images. Experimental results show that our model achieves state-of-the-art generation quality both quantitatively and qualitatively significantly outperforming both the prior finetuning-based and finetuning-free personalization baselines.
****
ConsistDreamer: 3D-Consistent 2D Diffusion for High-Fidelity Scene Editing
Jun-Kun Chen, Samuel Rota Bulò, Norman Müller, Lorenzo Porzi, Peter Kontschieder, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21071-21080

This paper proposes ConsistDreamer - a novel framework that lifts 2D diffusion models with 3D awareness and 3D consistency thus enabling high-fidelity instruction-guided scene editing. To overcome the fundamental limitation of missing 3D consistency in 2D diffusion models our key insight is to introduce three synergetic strategies that augment the input of the 2D diffusion model to become 3D-aware and to explicitly enforce 3D consistency during the training process. Specifically we design surrounding views as context-rich input for the 2D diffusion model and generate 3D-consistent structured noise instead of image-independent noise. Moreover we introduce self-supervised consistency-enforcing training within the per-scene editing procedure. Extensive evaluation shows that our ConsistDreamer achieves state-of-the-art performance for instruction-guided scene editing across various scenes and editing instructions particularly in complicated large-scale indoor scenes from ScanNet++ with significantly improved sharpness and fine-grained textures. Notably ConsistDreamer stands as the first work capable of successfully editing complex (e.g. plaid/checkered) patterns.
****
HandDiff: 3D Hand Pose Estimation with Diffusion on Image-Point Cloud
Wencan Cheng, Hao Tang, Luc Van Gool, Jong Hwan Ko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2274-2284
Extracting keypoint locations from input hand frames known as 3D hand pose estimation is a critical task in various human-computer interaction applications. Essentially the 3D hand pose estimation can be regarded as a 3D point subset generative problem conditioned on input frames. Thanks to the recent significant progress on diffusion-based generative models hand pose estimation can also benefit from the diffusion model to estimate keypoint locations with high quality. However directly deploying the existing diffusion models to solve hand pose estimation is non-trivial since they cannot achieve the complex permutation mapping and precise localization. Based on this motivation this paper proposes HandDiff a diffusion-based hand pose estimation model that iteratively denoises accurate hand pose conditioned on hand-shaped image-point clouds. In order to recover keypoint permutation and accurate location we further introduce joint-wise condition and local detail condition. Experimental results demonstrate that the proposed HandDiff significantly outperforms the existing approaches on four challenging hand pose benchmark datasets. Codes and pre-trained models are publicly available at https://github.com/cwc1260/HandDiff.
****
SNIDA: Unlocking Few-Shot Object Detection with Non-linear Semantic Decoupling Augmentation
Yanjie Wang, Xu Zou, Luxin Yan, Sheng Zhong, Jiahuan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12544-12553
Once only a few-shot annotated samples are available the performance of learning-based object detection would be heavily dropped. Many few-shot object detection (FSOD) methods have been proposed to tackle this issue by adopting image-level augmentations in linear manners. Nevertheless those handcrafted enhancements often suffer from limited diversity and lack of semantic awareness resulting in unsatisfactory performance. To this end we propose a Semantic-guided Non-linear Instance-level Data Augmentation method (SNIDA) for FSOD by decoupling the foreground and background to increase their diversities respectively. We design a semantic awareness enhancement strategy to separate objects from backgrounds. Concretely masks of instances are extracted by an unsupervised semantic segmentation module. Then the diversity of samples would be improved by fusing instances into different backgrounds. Considering the shortcomings of augmenting images in a limited transformation space of existing traditional data augmentation methods we introduce an object reconstruction enhancement module. The aim of this module is to generate sufficient diversity and non-linear training data at the instance level through a semantic-guided masked autoencoder. In this way the potential of data can be fully exploited in various object detection scenarios. Extensive experiments on PASCAL VOC and MS-COCO demonstrate that the proposed method outperform

s baselines by a large margin and achieves new state-of-the-art results under di
fferent shot settings.
****
On the Robustness of Large Multimodal Models Against Image Adversarial Attacks
Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, Ser-Nam Lim; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 24625-24634
Recent advances in instruction tuning have led to the development of State-of-th
e-Art Large Multimodal Models (LMMs). Given the novelty of these models the impa
ct of visual adversarial attacks on LMMs has not been thoroughly examined. We co
nduct a comprehensive study of the robustness of various LMMs against different
adversarial attacks evaluated across tasks including image classification image
captioning and Visual Question Answer (VQA). We find that in general LMMs are no
t robust to visual adversarial inputs. However our findings suggest that context
 provided to the model via prompts--such as questions in a QA pair--helps to mit
igate the effects of visual adversarial inputs. Notably the LMMs evaluated demon
strated remarkable resilience to such attacks on the ScienceQA task with only an
 8.10% drop in performance compared to their visual counterparts which dropped 9
9.73%. We also propose a new approach to real-world image classification which w
e term query decomposition. By incorporating existence queries into our input pr
ompt we observe diminished attack effectiveness and improvements in image classi
fication accuracy. This research highlights a previously under explored facet of
 LMM robustness and sets the stage for future work aimed at strengthening the re
silience of multimodal systems in adversarial environments.
****
SoundingActions: Learning How Actions Sound from Narrated Egocentric Videos
Changan Chen, Kumar Ashutosh, Rohit Girdhar, David Harwath, Kristen Grauman; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2024, pp. 27252-27262
We propose a novel self-supervised embedding to learn how actions sound from nar
rated in-the-wild egocentric videos. Whereas existing methods rely on curated da
ta with known audio-visual correspondence our multimodal contrastive-consensus c
oding (MC3) embedding reinforces the associations between audio language and vis
ion when all modality pairs agree while diminishing those associations when any
one pair does not. We show our approach can successfully discover how the long t
ail of human actions sound from egocentric video outperforming an array of recen
t multimodal embedding techniques on two datasets (Ego4D and EPIC-Sounds) and mu
ltiple cross-modal tasks.
****
Not All Voxels Are Equal: Hardness-Aware Semantic Scene Completion with Self-Dis
tillation
Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, Jianke Zhu;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2024, pp. 14792-14801
Semantic scene completion also known as semantic occupancy prediction can provid
e dense geometric and semantic information for autonomous vehicles which attract
s the increasing attention of both academia and industry. Unfortunately existing
 methods usually formulate this task as a voxel-wise classification problem and
treat each voxel equally in 3D space during training. As the hard voxels have no
t been paid enough attention the performance in some challenging regions is limi
ted. The 3D dense space typically contains a large number of empty voxels which
are easy to learn but require amounts of computation due to handling all the vox
els uniformly for the existing models. Furthermore the voxels in the boundary re
gion are more challenging to differentiate than those in the interior. In this p
aper we propose HASSC approach to train the semantic scene completion model with
 hardness-aware design. The global hardness from the network optimization proces
s is defined for dynamical hard voxel selection. Then the local hardness with ge
ometric anisotropy is adopted for voxel-wise refinement. Besides self-distillati
on strategy is introduced to make training process stable and consistent. Extens
ive experiments show that our HASSC scheme can effectively promote the accuracy

of the baseline model without incurring the extra inference cost. Source code is available at: https://github.com/songw-zju/HASSC.
****
3D-LFM: Lifting Foundation Model
Mosam Dabhi, László A. Jeni, Simon Lucey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10466-10475
The lifting of a 3D structure and camera from 2D landmarks is at the cornerstone of the discipline of computer vision. Traditional methods have been confined to specific rigid objects such as those in Perspective-n-Point (PnP) problems but deep learning has expanded our capability to reconstruct a wide range of object classes (e.g. C3DPO [??] and PAUL [??]) with resilience to noise occlusions and perspective distortions. However all these techniques have been limited by the fundamental need to establish correspondences across the 3D training data significantly limiting their utility to applications where one has an abundance of "in-correspondence" 3D data. Our approach harnesses the inherent permutation equivariance of transformers to manage varying numbers of points per 3D data instance withstands occlusions and generalizes to unseen categories. We demonstrate state-of-the-art performance across 2D-3D lifting task benchmarks. Since our approach can be trained across such a broad class of structures we refer to it simply as a 3D Lifting Foundation Model (3D-LFM) -- the first of its kind.
****
VP3D: Unleashing 2D Visual Prompt for Text-to-3D Generation
Yang Chen, Yingwei Pan, Haibo Yang, Ting Yao, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4896-4905
Recent innovations on text-to-3D generation have featured Score Distillation Sampling (SDS) which enables the zero-shot learning of implicit 3D models (NeRF) by directly distilling prior knowledge from 2D diffusion models. However current SDS-based models still struggle with intricate text prompts and commonly result in distorted 3D models with unrealistic textures or cross-view inconsistency issues. In this work we introduce a novel Visual Prompt-guided text-to-3D diffusion model (VP3D) that explicitly unleashes the visual appearance knowledge in 2D visual prompt to boost text-to-3D generation. Instead of solely supervising SDS with text prompt VP3D first capitalizes on 2D diffusion model to generate a high-quality image from input text which subsequently acts as visual prompt to strengthen SDS optimization with explicit visual appearance. Meanwhile we couple the SDS optimization with additional differentiable reward function that encourages rendering images of 3D models to better visually align with 2D visual prompt and semantically match with text prompt. Through extensive experiments we show that the 2D Visual Prompt in our VP3D significantly eases the learning of visual appearance of 3D models and thus leads to higher visual fidelity with more detailed textures. It is also appealing in view that when replacing the self-generating visual prompt with a given reference image VP3D is able to trigger a new task of stylized text-to-3D generation. Our project page is available at https://vp3d-cvpr24.github.io.
****
MonoHair: High-Fidelity Hair Modeling from a Monocular Video
Keyu Wu, Lingchen Yang, Zhiyi Kuang, Yao Feng, Xutao Han, Yuefan Shen, Hongbo Fu, Kun Zhou, Youyi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24164-24173
Undoubtedly high-fidelity 3D hair is crucial for achieving realism artistic expression and immersion in computer graphics. While existing 3D hair modeling methods have achieved impressive performance the challenge of achieving high-quality hair reconstruction persists: they either require strict capture conditions making practical applications difficult or heavily rely on learned prior data obscuring fine-grained details in images. To address these challenges we propose MonoHair a generic framework to achieve high-fidelity hair reconstruction from a monocular video without specific requirements for environments. Our approach bifurcates the hair modeling process into two main stages: precise exterior reconstruction and interior structure inference. The exterior is meticulously crafted using

our Patch-based Multi-View Optimization PMVO. This method strategically collects and integrates hair information from multiple views independent of prior data to produce a high-fidelity exterior 3D line map. This map not only captures intricate details but also facilitates the inference of the hair's inner structure. For the interior we employ a data-driven multi-view 3D hair reconstruction method. This method utilizes 2D structural renderings derived from the reconstructed exterior mirroring the synthetic 2D inputs used during training. This alignment effectively bridges the domain gap between our training data and real-world data thereby enhancing the accuracy and reliability of our interior structure inference. Lastly we generate a strand model and resolve the directional ambiguity by our hair growth algorithm. Our experiments demonstrate that our method exhibits robustness across diverse hairstyles and achieves state-of-the-art performance. For more results please refer to our project page https://keyuwu-cs.github.io/MonoHair/
****
Content-Style Decoupling for Unsupervised Makeup Transfer without Generating Pseudo Ground Truth
Zhaoyang Sun, Shengwu Xiong, Yaxiong Chen, Yi Rong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7601-7610
The absence of real targets to guide the model training is one of the main problems with the makeup transfer task. Most existing methods tackle this problem by synthesizing pseudo ground truths (PGTs). However the generated PGTs are often sub-optimal and their imprecision will eventually lead to performance degradation. To alleviate this issue in this paper we propose a novel Content-Style Decoupled Makeup Transfer (CSD-MT) method which works in a purely unsupervised manner and thus eliminates the negative effects of generating PGTs. Specifically based on the frequency characteristics analysis we assume that the low-frequency (LF) component of a face image is more associated with its makeup style information while the high-frequency (HF) component is more related to its content details. This assumption allows CSD-MT to decouple the content and makeup style information in each face image through the frequency decomposition. After that CSD-MT realizes makeup transfer by maximizing the consistency of these two types of information between the transferred result and input images respectively. Two newly designed loss functions are also introduced to further improve the transfer performance. Extensive quantitative and qualitative analyses show the effectiveness of our CSD-MT method. Our code is available at https://github.com/Snowfallingplum/CSD-MT.
****
One Prompt Word is Enough to Boost Adversarial Robustness for Pre-trained Vision-Language Models
Lin Li, Haoyan Guan, Jianing Qiu, Michael Spratling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24408-24419
Large pre-trained Vision-Language Models (VLMs) like CLIP despite having remarkable generalization ability are highly vulnerable to adversarial examples. This work studies the adversarial robustness of VLMs from the novel perspective of the text prompt instead of the extensively studied model weights (frozen in this work). We first show that the effectiveness of both adversarial attack and defense are sensitive to the used text prompt. Inspired by this we propose a method to improve resilience to adversarial attacks by learning a robust text prompt for VLMs. The proposed method named Adversarial Prompt Tuning (APT) is effective while being both computationally and data efficient. Extensive experiments are conducted across 15 datasets and 4 data sparsity schemes (from 1-shot to full training data settings) to show APT's superiority over hand-engineered prompts and other state-of-the-art adaption methods. APT demonstrated excellent abilities in terms of the in-distribution performance and the generalization under input distribution shift and across datasets. Surprisingly by simply adding one learned word to the prompts APT can significantly boost the accuracy and robustness (epsilon=4/255) over the hand-engineered prompts by +13% and +8.5% on average respectivel

y. The improvement further increases in our most effective setting to +26.4% for accuracy and +16.7% for robustness. Code is available at https://github.com/TreeLLi/APT.
****
A Versatile Framework for Continual Test-Time Domain Adaptation: Balancing Discriminability and Generalizability
Xu Yang, Xuan Chen, Moqi Li, Kun Wei, Cheng Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23731-23740
Continual test-time domain adaptation (CTTA) aims to adapt the source pre-trained model to a continually changing target domain without additional data acquisition or labeling costs. This issue necessitates an initial performance enhancement within the present domain without labels while concurrently averting an excessive bias toward the current domain. Such bias exacerbates catastrophic forgetting and diminishes the generalization ability to future domains. To tackle the problem this paper designs a versatile framework to capture high-quality supervision signals from three aspects: 1) The adaptive thresholds are employed to determine the reliability of pseudo-labels; 2) The knowledge from the source pre-trained model is utilized to adjust the unreliable one and 3) By evaluating past supervision signals we calculate a diversity score to ensure subsequent generalization. In this way we form a complete supervisory signal generation framework which can capture the current domain discriminative and reserve generalization in future domains. Finally to avoid catastrophic forgetting we design a weighted soft parameter alignment method to explore the knowledge from the source model. Extensive experimental results demonstrate that our method performs well on several benchmark datasets.
****
Quantifying Uncertainty in Motion Prediction with Variational Bayesian Mixture
Juanwu Lu, Can Cui, Yunsheng Ma, Aniket Bera, Ziran Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15428-15437
Safety and robustness are crucial factors in developing trustworthy autonomous vehicles. One essential aspect of addressing these factors is to equip vehicles with the capability to predict future trajectories for all moving objects in the surroundings and quantify prediction uncertainties. In this paper we propose the Sequential Neural Variational Agent (SeNeVA) a generative model that describes the distribution of future trajectories for a single moving object. Our approach can distinguish Out-of-Distribution data while quantifying uncertainty and achieving competitive performance compared to state-of-the-art methods on the Argoverse 2 and INTERACTION datasets. Specifically a 0.446 meters minimum Final Displacement Error a 0.203 meters minimum Average Displacement Error and a 5.35% Miss Rate are achieved on the INTERACTION test set. Extensive qualitative and quantitative analysis is also provided to evaluate the proposed model. Our open-source code is available at https://github.com/PurdueDigitalTwin/seneva.
****
You Only Need Less Attention at Each Stage in Vision Transformers
Shuoxi Zhang, Hanpeng Liu, Stephen Lin, Kun He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6057-6066
The advent of Vision Transformers (ViTs) marks a substantial paradigm shift in the realm of computer vision. ViTs capture the global information of images through self-attention modules which perform dot product computations among patchified image tokens. While self-attention modules empower ViTs to capture long-range dependencies the computational complexity grows quadratically with the number of tokens which is a major hindrance to the practical application of ViTs. Moreover the self-attention mechanism in deep ViTs is also susceptible to the attention saturation issue. Accordingly we argue against the necessity of computing the attention scores in every layer and we propose the Less-Attention Vision Transformer (LaViT) which computes only a few attention operations at each stage and calculates the subsequent feature alignments in other layers via attention transformations that leverage the previously calculated attention scores. This novel app

roach can mitigate two primary issues plaguing traditional self-attention modules: the heavy computational burden and attention saturation. Our proposed architecture offers superior efficiency and ease of implementation merely requiring matrix multiplications that are highly optimized in contemporary deep learning frameworks. Moreover our architecture demonstrates exceptional performance across various vision tasks including classification detection and segmentation.
****
Sieve: Multimodal Dataset Pruning using Image Captioning Models
Anas Mahmoud, Mostafa Elhoushi, Amro Abbas, Yu Yang, Newsha Ardalani, Hugh Leather, Ari S. Morcos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22423-22432
Vision-Language Models (VLMs) are pretrained on large diverse and noisy web-crawled datasets. This underscores the critical need for dataset pruning as the quality of these datasets is strongly correlated with the performance of VLMs on downstream tasks. Using CLIPScore from a pretrained model to only train models using highly-aligned samples is one of the most successful methods for pruning. We argue that this approach suffers from multiple limitations including: false positives and negatives due to CLIP's pretraining on noisy labels. We propose a pruning signal Sieve that employs synthetic captions generated by image-captioning models pretrained on small diverse and well-aligned image-text pairs to evaluate the alignment of noisy image-text pairs. To bridge the gap between the limited diversity of generated captions and the high diversity of alternative text (alt-text) we estimate the semantic textual similarity in the embedding space of a language model pretrained on unlabeled text corpus. Using DataComp a multimodal dataset filtering benchmark when evaluating on 38 downstream tasks our pruning approach surpasses CLIPScore by 2.6% and 1.7% on medium and large scale respectively. In addition on retrieval tasks Sieve leads to a significant improvement of 2.7% and 4.5% on medium and large scale respectively.
****
Generalizable Novel-View Synthesis using a Stereo Camera
Haechan Lee, Wonjoon Jin, Seung-Hwan Baek, Sunghyun Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4939-4948
In this paper we propose the first generalizable view synthesis approach that specifically targets multi-view stereo-camera images. Since recent stereo matching has demonstrated accurate geometry prediction we introduce stereo matching into novel-view synthesis for high-quality geometry reconstruction. To this end this paper proposes a novel framework dubbed StereoNeRF which integrates stereo matching into a NeRF-based generalizable view synthesis approach. StereoNeRF is equipped with three key components to effectively exploit stereo matching in novel-view synthesis: a stereo feature extractor a depth-guided plane-sweeping and a stereo depth loss. Moreover we propose the StereoNVS dataset the first multi-view dataset of stereo-camera images encompassing a wide variety of both real and synthetic scenes. Our experimental results demonstrate that StereoNeRF surpasses previous approaches in generalizable view synthesis.
****
Dynamic LiDAR Re-simulation using Compositional Neural Fields
Hanfeng Wu, Xingxing Zuo, Stefan Leutenegger, Or Litany, Konrad Schindler, Shengyu Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19988-19998
We introduce DyNFL a novel neural field-based approach for high-fidelity re-simulation of LiDAR scans in dynamic driving scenes. DyNFL processes LiDAR measurements from dynamic environments accompanied by bounding boxes of moving objects to construct an editable neural field. This field comprising separately reconstructed static background and dynamic objects allows users to modify viewpoints adjust object positions and seamlessly add or remove objects in the re-simulated scene. A key innovation of our method is the neural field composition technique which effectively integrates reconstructed neural assets from various scenes through a ray drop test accounting for occlusions and transparent surfaces. Our evaluation with both synthetic and real-world environments demonstrates that DyNFL sub

stantially improves dynamic scene LiDAR simulation offering a combination of physical fidelity and flexible editing capabilities. Project page: https://shengyuh.github.io/dynfl
****
Explaining CLIP's Performance Disparities on Data from Blind/Low Vision Users
Daniela Massiceti, Camilla Longden, Agnieszka Slowik, Samuel Wills, Martin Grayson, Cecily Morrison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12172-12182
Large multi-modal models (LMMs) hold the potential to usher in a new era of automated visual assistance for people who are blind or low vision (BLV). Yet these models have not been systematically evaluated on data captured by BLV users. We address this by empirically assessing CLIP a widely-used LMM likely to underpin many assistive technologies. Testing 25 CLIP variants in a zero-shot classification task we find that their accuracy is 15 percentage points lower on average for images captured by BLV users than web-crawled images. This disparity stems from CLIP's sensitivities to 1) image content (e.g. not recognizing disability objects as well as other objects); 2) image quality (e.g. not being robust to lighting variation); and 3) text content (e.g. not recognizing objects described by tactile adjectives as well as visual ones). We delve deeper with a textual analysis of three common pre-training datasets: LAION-400M LAION-2B and DataComp-1B showing that disability content is rarely mentioned. We then provide three examples that illustrate how the performance disparities extend to three downstream models underpinned by CLIP: OWL-ViT CLIPSeg and DALL-E2. We find that few-shot learning with as few as 5 images can mitigate CLIP's quality-of-service disparities for BLV users in some scenarios which we discuss alongside a set of other possible mitigations.
****
AETTA: Label-Free Accuracy Estimation for Test-Time Adaptation
Taeckyung Lee, Sorn Chottananurak, Taesik Gong, Sung-Ju Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28643-28652
Test-time adaptation (TTA) has emerged as a viable solution to adapt pre-trained models to domain shifts using unlabeled test data. However TTA faces challenges of adaptation failures due to its reliance on blind adaptation to unknown test samples in dynamic scenarios. Traditional methods for out-of-distribution performance estimation are limited by unrealistic assumptions in the TTA context such as requiring labeled data or re-training models. To address this issue we propose AETTA a label-free accuracy estimation algorithm for TTA. We propose the prediction disagreement as the accuracy estimate calculated by comparing the target model prediction with dropout inferences. We then improve the prediction disagreement to extend the applicability of AETTA under adaptation failures. Our extensive evaluation with four baselines and six TTA methods demonstrates that AETTA shows an average of 19.8%p more accurate estimation compared with the baselines. We further demonstrate the effectiveness of accuracy estimation with a model recovery case study showcasing the practicality of our model recovery based on accuracy estimation. The source code is available at https://github.com/taeckyung/AETTA.
****
Digital Life Project: Autonomous 3D Characters with Social Intelligence
Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, Liang Pan, Xiangyu Fan, Han Du, Peng Gao, Zhitao Yang, Yang Gao, Jiaqi Li, Tianxiang Ren, Yukun Wei, Xiaogang Wang, Chen Change Loy, Lei Yang, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 582-592
In this work we present Digital Life Project a framework utilizing language as the universal medium to build autonomous 3D characters who are capable of engaging in social interactions and expressing with articulated body motions thereby simulating life in a digital environment. Our framework comprises two primary components: 1) SocioMind: a meticulously crafted digital brain that models personalities with systematic few-shot exemplars incorporates a reflection process based

on psychology principles and emulates autonomy by initiating dialogue topics; 2) MoMat-MoGen: a text-driven motion synthesis paradigm for controlling the character's digital body. It integrates motion matching a proven industry technique to ensure motion quality with cutting-edge advancements in motion generation for diversity. Extensive experiments demonstrate that each module achieves state-of-the-art performance in its respective domain. Collectively they enable virtual characters to initiate and sustain dialogues autonomously while evolving their socio-psychological states. Concurrently these characters can perform contextually relevant bodily movements. Additionally an extension of DLP enables a virtual character to recognize and appropriately respond to human players' actions.
****
An Empirical Study of the Generalization Ability of Lidar 3D Object Detectors to Unseen Domains
George Eskandar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23815-23825
3D Object Detectors (3D-OD) are crucial for understanding the environment in many robotic tasks especially autonomous driving. Including 3D information via Lidar sensors improves accuracy greatly. However such detectors perform poorly on domains they were not trained on i.e. different locations sensors weather etc. limiting their reliability in safety-critical applications. There exist methods to adapt 3D-ODs to these domains; however these methods treat 3D-ODs as a black box neglecting underlying architectural decisions and source-domain training strategies. Instead we dive deep into the details of 3D-ODs focusing our efforts on fundamental factors that influence robustness prior to domain adaptation. We systematically investigate four design choices (and the interplay between them) often overlooked in 3D-OD robustness and domain adaptation: architecture voxel encoding data augmentations and anchor strategies. We assess their impact on the robustness of nine state-of-the-art 3D-ODs across six benchmarks encompassing three types of domain gaps - sensor type weather and location. Our main findings are: (1) transformer backbones with local point features are more robust than 3D CNNs (2) test-time anchor size adjustment is crucial for adaptation across geographical locations significantly boosting scores without retraining (3) source-domain augmentations allow the model to generalize to low-resolution sensors and (4) surprisingly robustness to bad weather is improved when training directly on more clean weather data than on training with bad weather data. We outline our main conclusions and findings to provide practical guidance on developing more robust 3D-ODs.
****
Unsupervised Universal Image Segmentation
Dantong Niu, Xudong Wang, Xinyang Han, Long Lian, Roei Herzig, Trevor Darrell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22744-22754
Several unsupervised image segmentation approaches have been proposed which eliminate the need for dense manually-annotated segmentation masks; current models separately handle either semantic segmentation (e.g. STEGO) or class-agnostic instance segmentation (e.g. CutLER) but not both (i.e. panoptic segmentation). We propose an Unsupervised Universal Segmentation model (U2Seg) adept at performing various image segmentation tasks---instance semantic and panoptic---using a novel unified framework. U2Seg generates pseudo semantic labels for these segmentation tasks via leveraging self-supervised models followed by clustering; each cluster represents different semantic and/or instance membership of pixels. We then self-train the model on these pseudo semantic labels yielding substantial performance gains over specialized methods tailored to each task: a +2.6 APbox boost (vs. CutLER) in unsupervised instance segmentation on COCO and a +7.0 PixelAcc increase (vs. STEGO) in unsupervised semantic segmentation on COCOStuff. Moreover our method sets up a new baseline for unsupervised panoptic segmentation which has not been previously explored. U2Seg is also a strong pretrained model for few-shot segmentation surpassing CutLER by +5.0 APmask when trained on a low-data regime e.g. only 1% COCO labels. We hope our simple yet effective method can inspire more research on unsupervised universal image segmentation.

****
Rethinking Prior Information Generation with CLIP for Few-Shot Segmentation
Jin Wang, Bingfeng Zhang, Jian Pang, Honglong Chen, Weifeng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3941-3951
Few-shot segmentation remains challenging due to the limitations of its labeling information for unseen classes. Most previous approaches rely on extracting high-level feature maps from the frozen visual encoder to compute the pixel-wise similarity as a key prior guidance for the decoder. However such a prior representation suffers from coarse granularity and poor generalization to new classes since these high-level feature maps have obvious category bias. In this work we propose to replace the visual prior representation with the visual-text alignment capacity to capture more reliable guidance and enhance the model generalization. Specifically we design two kinds of training-free prior information generation strategy that attempts to utilize the semantic alignment capability of the Contrastive Language-Image Pre-training model (CLIP) to locate the target class. Besides to acquire more accurate prior guidance we build a high-order relationship of attention maps and utilize it to refine the initial prior information. Experiments on both the PASCAL-5i and COCO-20i datasets show that our method obtains a clearly substantial improvement and reaches the new state-of-the-art performance. The code is available on the project website.
****
SingularTrajectory: Universal Trajectory Predictor Using Diffusion Model
Inhwan Bae, Young-Jae Park, Hae-Gon Jeon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17890-17901
There are five types of trajectory prediction tasks: deterministic stochastic domain adaptation momentary observation and few-shot. These associated tasks are defined by various factors such as the length of input paths data split and pre-processing methods. Interestingly even though they commonly take sequential coordinates of observations as input and infer future paths in the same coordinates as output designing specialized architectures for each task is still necessary. For the other task generality issues can lead to sub-optimal performances. In this paper we propose SingularTrajectory a diffusion-based universal trajectory prediction framework to reduce the performance gap across the five tasks. The core of SingularTrajectory is to unify a variety of human dynamics representations on the associated tasks. To do this we first build a Singular space to project all types of motion patterns from each task into one embedding space. We next propose an adaptive anchor working in the Singular space. Unlike traditional fixed anchor methods that sometimes yield unacceptable paths our adaptive anchor enables correct anchors which are put into a wrong location based on a traversability map. Finally we adopt a diffusion-based predictor to further enhance the prototype paths using a cascaded denoising process. Our unified framework ensures the generality across various benchmark settings such as input modality and trajectory lengths. Extensive experiments on five public benchmarks demonstrate that SingularTrajectory substantially outperforms existing models highlighting its effectiveness in estimating general dynamics of human movements. Code is publicly available at https://github.com/inhwanbae/SingularTrajectory.
****
Generating Handwritten Mathematical Expressions From Symbol Graphs: An End-to-End Pipeline
Yu Chen, Fei Gao, Yanguang Zhang, Maoying Qiao, Nannan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15675-15685
In this paper we explore a novel challenging generation task i.e. Handwritten Mathematical Expression Generation (HMEG) from symbolic sequences. Since symbolic sequences are naturally graph-structured data we formulate HMEG as a graph-to-image (G2I) generation problem. Unlike the generation of natural images HMEG requires critic layout clarity for synthesizing correct and recognizable formulas but has no real masks available to supervise the learning process. To alleviate this challenge we propose a novel end-to-end G2I generation pipeline (i.e. graph -

layout - mask - image) which requires no real masks or nondifferentiable alignme
nt between layouts and masks. Technically to boost the capacity of predicting de
tailed relations among adjacent symbols we propose a Less-is-More (LiM) learning
 strategy. In addition we design a differentiable layout refinement module which
 maps bounding boxes to pixel-level soft masks so as to further alleviate ambigu
ous layout areas. Our whole model including layout prediction mask refinement an
d image generation can be jointly optimized in an end-to-end manner. Experimenta
l results show that our model can generate high-quality HME images and outperfor
ms previous generative methods. Besides a series of ablations study demonstrate
effectiveness of the proposed techniques. Finally we validate that our generated
 images promisingly boosts the performance of HME recognition models through dat
a augmentation. Our code and results are available at: https://github.com/AiArt-
HDU/HMEG.
****
A Closer Look at the Few-Shot Adaptation of Large Vision-Language Models
Julio Silva-Rodríguez, Sina Hajimiri, Ismail Ben Ayed, Jose Dolz; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 23681-23690
Efficient transfer learning (ETL) is receiving increasing attention to adapt lar
ge pre-trained language-vision models on downstream tasks with a few labeled sam
ples. While significant progress has been made we reveal that state-of-the-art E
TL approaches exhibit strong performance only in narrowly-defined experimental s
etups and with a careful adjustment of hyperparameters based on a large corpus o
f labeled samples. In particular we make two interesting and surprising empirica
l observations. First to outperform a simple Linear Probing baseline these metho
ds require to optimize their hyper-parameters on each target task. And second th
ey typically underperform --sometimes dramatically-- standard zero-shot predicti
ons in the presence of distributional drifts. Motivated by the unrealistic assum
ptions made in the existing literature i.e. access to a large validation set and
 case-specific grid-search for optimal hyperparameters we propose a novel approa
ch that meets the requirements of real-world scenarios. More concretely we intro
duce a CLass-Adaptive linear Probe (CLAP) objective whose balancing term is opti
mized via an adaptation of the general Augmented Lagrangian method tailored to t
his context. We comprehensively evaluate CLAP on a broad span of datasets and sc
enarios demonstrating that it consistently outperforms SoTA approaches while yet
 being a much more efficient alternative.
****
Generative Rendering: Controllable 4D-Guided Video Generation with 2D Diffusion
Models
Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Yang W
ang, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2024, pp. 7611-7620
Traditional 3D content creation tools empower users to bring their imagination t
o life by giving them direct control over a scene's geometry appearance motion a
nd camera path. Creating computer-generated videos however is a tedious manual p
rocess which can be automated by emerging text-to-video diffusion models. Despit
e great promise video diffusion models are difficult to control hindering users
to apply their creativity rather than amplifying it. To address this challenge w
e present a novel approach that combines the controllability of dynamic 3D meshe
s with the expressivity and editability of emerging diffusion models. For this p
urpose our approach takes an animated low-fidelity rendered mesh as input and in
jects the ground truth correspondence information obtained from the dynamic mesh
 into various stages of a pre-trained text-to-image generation model to output h
igh-quality and temporally consistent frames. We demonstrate our approach on var
ious examples where motion can be obtained by animating rigged assets or changin
g the camera path.
****
Relightable Gaussian Codec Avatars
Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, Giljoo Nam; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

2024, pp. 130-141
The fidelity of relighting is bounded by both geometry and appearance representations. For geometry both mesh and volumetric approaches have difficulty modeling intricate structures like 3D hair geometry. For appearance existing relighting models are limited in fidelity and often too slow to render in real-time with high-resolution continuous environments. In this work we present Relightable Gaussian Codec Avatars a method to build high-fidelity relightable head avatars that can be animated to generate novel expressions. Our geometry model based on 3D Gaussians can capture 3D-consistent sub-millimeter details such as hair strands and pores on dynamic face sequences. To support diverse materials of human heads such as the eyes skin and hair in a unified manner we present a novel relightable appearance model based on learnable radiance transfer. Together with global illumination-aware spherical harmonics for the diffuse components we achieve real-time relighting with all-frequency reflections using spherical Gaussians. This appearance model can be efficiently relit under both point light and continuous illumination. We further improve the fidelity of eye reflections and enable explicit gaze control by introducing relightable explicit eye models. Our method outperforms existing approaches without compromising real-time performance. We also demonstrate real-time relighting of avatars on a tethered consumer VR headset showcasing the efficiency and fidelity of our avatars.
****
Why Not Use Your Textbook? Knowledge-Enhanced Procedure Planning of Instructional Videos
Kumaranage Ravindu Yasas Nagasinghe, Honglu Zhou, Malitha Gunawardhana, Martin Renqiang Min, Daniel Harari, Muhammad Haris Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18816-18826
In this paper we explore the capability of an agent to construct a logical sequence of action steps thereby assembling a strategic procedural plan. This plan is crucial for navigating from an initial visual observation to a target visual outcome as depicted in real-life instructional videos. Existing works have attained partial success by extensively leveraging various sources of information available in the datasets such as heavy intermediate visual observations procedural names or natural language step-by-step instructions for features or supervision signals. However the task remains formidable due to the implicit causal constraints in the sequencing of steps and the variability inherent in multiple feasible plans. To tackle these intricacies that previous efforts have overlooked we propose to enhance the agent's capabilities by infusing it with procedural knowledge. This knowledge sourced from training procedure plans and structured as a directed weighted graph equips the agent to better navigate the complexities of step sequencing and its potential variations. We coin our approach KEPP a novel Knowledge-Enhanced Procedure Planning system which harnesses a probabilistic procedural knowledge graph extracted from training data effectively acting as a comprehensive textbook for the training domain. Experimental evaluations across three widely-used datasets under settings of varying complexity reveal that KEPP attains superior state-of-the-art results while requiring only minimal supervision. Code and trained model are available at https://github.com/Ravindu-Yasas-Nagasinghe/KEPP
****
Global and Hierarchical Geometry Consistency Priors for Few-shot NeRFs in Indoor Scenes
Xiaotian Sun, Qingshan Xu, Xinjie Yang, Yu Zang, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20530-20539
It is challenging for Neural Radiance Fields (NeRFs) in the few-shot setting to reconstruct high-quality novel views and depth maps in 360^\circ outward-facing indoor scenes. The captured sparse views for these scenes usually contain large viewpoint variations. This greatly reduces the potential consistency between views leading NeRFs to degrade a lot in these scenarios. Existing methods usually leverage pretrained depth prediction models to improve NeRFs. However these methods cannot guarantee geometry consistency due to the inherent geometry ambiguity

in the pretrained models thus limiting NeRFs' performance. In this work we prese
nt P\textsuperscript 2 NeRF to capture global and hierarchical geometry consiste
ncy priors from pretrained models thus facilitating few-shot NeRFs in 360^\circ
outward-facing indoor scenes. On the one hand we propose a matching-based geomet
ry warm-up strategy to provide global geometry consistency priors for NeRFs. Thi
s effectively avoids the overfitting of early training with sparse inputs. On th
e other hand we propose a group depth ranking loss and ray weight mask regulariz
ation based on the monocular depth estimation model. This provides hierarchical
geometry consistency priors for NeRFs. As a result our approach can fully levera
ge the geometry consistency priors from pretrained models and help few-shot NeRF
s achieve state-of-the-art performance on two challenging indoor datasets. Our c
ode is released at https://github.com/XT5un/P2NeRF.
****
FreeKD: Knowledge Distillation via Semantic Frequency Prompt
Yuan Zhang, Tao Huang, Jiaming Liu, Tao Jiang, Kuan Cheng, Shanghang Zhang; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 15931-15940
Knowledge distillation (KD) has been applied to various tasks successfully and m
ainstream methods typically boost the student model via spatial imitation losses
. However the consecutive downsamplings induced in the spatial domain of teacher
 model is a type of corruption hindering the student from analyzing what specifi
c information needs to be imitated which results in accuracy degradation. To bet
ter understand the underlying pattern of corrupted feature maps we shift our att
ention to the frequency domain. During frequency distillation we encounter a new
 challenge: the low-frequency bands convey general but minimal context while the
 high are more informative but also introduce noise. Not each pixel within the f
requency bands contributes equally to the performance. To address the above prob
lem: (1) We propose the Frequency Prompt plugged into the teacher model absorbin
g the semantic frequency context during finetuning. (2) During the distillation
period a pixel-wise frequency mask is generated via Frequency Prompt to localize
 those pixel of interests (PoIs) in various frequency bands. Additionally we emp
loy a position-aware relational frequency loss for dense prediction tasks delive
ring a high-order spatial enhancement to the student model. We dub our Frequency
 Knowledge Distillation method as FreeKD which determines the optimal localizati
on and extent for the frequency distillation. Extensive experiments demonstrate
that FreeKD not only outperforms spatial-based distillation methods consistently
 on dense prediction tasks (e.g. FreeKD brings 3.8 AP gains for RepPoints-R50 on
 COCO2017 and 4.55 mIoU gains for PSPNet-R18 on Cityscapes) but also conveys mor
e robustness to the student. Notably we also validate the generalization of our
approach on large-scale vision models (e.g. DINO and SAM).
****
Can't Make an Omelette Without Breaking Some Eggs: Plausible Action Anticipation
 Using Large Video-Language Models
Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, Kwonjoon Lee; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1
8580-18590
We introduce PlausiVL a large video-language model for anticipating action seque
nces that are plausible in the real-world. While significant efforts have been m
ade towards anticipating future actions prior approaches do not take into accoun
t the aspect of plausibility in an action sequence. To address this limitation w
e explore the generative capability of a large video-language model in our work
and further develop the understanding of plausibility in an action sequence by i
ntroducing two objective functions a counterfactual-based plausible action seque
nce learning loss and a long-horizon action repetition loss. We utilize temporal
 logical constraints as well as verb-noun action pair logical constraints to cre
ate implausible/counterfactual action sequences and use them to train the model
with plausible action sequence learning loss. This loss helps the model to diffe
rentiate between plausible and not plausible action sequences and also helps the
 model to learn implicit temporal cues crucial for the task of action anticipati
on. The long-horizon action repetition loss puts a higher penalty on the actions

that are more prone to repetition over a longer temporal window. With this penalization the model is able to generate diverse plausible action sequences. We evaluate our approach on two large-scale datasets Ego4D and EPIC-Kitchens-100 and show improvements on the task of action anticipation.
****
On the Estimation of Image-matching Uncertainty in Visual Place Recognition
Mubariz Zaffar, Liangliang Nan, Julian F. P. Kooij; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17743-17753
In Visual Place Recognition (VPR) the pose of a query image is estimated by comparing the image to a map of reference images with known reference poses. As is typical for image retrieval problems a feature extractor maps the query and reference images to a feature space where a nearest neighbor search is then performed. However till recently little attention has been given to quantifying the confidence that a retrieved reference image is a correct match. Highly certain but incorrect retrieval can lead to catastrophic failure of VPR-based localization pipelines. This work compares for the first time the main approaches for estimating the image-matching uncertainty including the traditional retrieval-based uncertainty estimation more recent data-driven aleatoric uncertainty estimation and the compute-intensive geometric verification. We further formulate a simple baseline method "SUE" which unlike the other methods considers the freely-available poses of the reference images in the map. Our experiments reveal that a simple L2-distance between the query and reference descriptors is already a better estimate of image-matching uncertainty than current data-driven approaches. SUE outperforms the other efficient uncertainty estimation methods and its uncertainty estimates complement the computationally expensive geometric verification approach. Future works for uncertainty estimation in VPR should consider the baselines discussed in this work.
****
Mask Grounding for Referring Image Segmentation
Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, Gao Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26573-26583
Referring Image Segmentation (RIS) is a challenging task that requires an algorithm to segment objects referred by free-form language expressions. Despite significant progress in recent years most state-of-the-art (SOTA) methods still suffer from considerable language-image modality gap at the pixel and word level. These methods generally 1) rely on sentence-level language features for language-image alignment and 2) lack explicit training supervision for fine-grained visual grounding. Consequently they exhibit weak object-level correspondence between visual and language features. Without well-grounded features prior methods struggle to understand complex expressions that require strong reasoning over relationships among multiple objects especially when dealing with rarely used or ambiguous clauses. To tackle this challenge we introduce a novel Mask Grounding auxiliary task that significantly improves visual grounding within language features by explicitly teaching the model to learn fine-grained correspondence between masked textual tokens and their matching visual objects. Mask Grounding can be directly used on prior RIS methods and consistently bring improvements. Furthermore to holistically address the modality gap we also design a cross-modal alignment loss and an accompanying alignment module. These additions work synergistically with Mask Grounding. With all these techniques our comprehensive approach culminates in MagNet (Mask-grounded Network) an architecture that significantly outperforms prior arts on three key benchmarks (RefCOCO RefCOCO+ and G-Ref) demonstrating our method's effectiveness in addressing current limitations of RIS algorithms. Our code and pre-trained weights will be released.
****
Single-to-Dual-View Adaptation for Egocentric 3D Hand Pose Estimation
Ruicong Liu, Takehiko Ohkawa, Mingfang Zhang, Yoichi Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 677-686

The pursuit of accurate 3D hand pose estimation stands as a keystone for understanding human activity in the realm of egocentric vision. The majority of existing estimation methods still rely on single-view images as input leading to potential limitations e.g. limited field-of-view and ambiguity in depth. To address these problems adding another camera to better capture the shape of hands is a practical direction. However existing multi-view hand pose estimation methods suffer from two main drawbacks: 1) Requiring multi-view annotations for training which are expensive. 2) During testing the model becomes inapplicable if camera parameters/layout are not the same as those used in training. In this paper we propose a novel Single-to-Dual-view adaptation (S2DHand) solution that adapts a pre-trained single-view estimator to dual views. Compared with existing multi-view training methods 1) our adaptation process is unsupervised eliminating the need for multi-view annotation. 2) Moreover our method can handle arbitrary dual-view pairs with unknown camera parameters making the model applicable to diverse camera settings. Specifically S2DHand is built on certain stereo constraints including pair-wise cross-view consensus and invariance of transformation between both views. These two stereo constraints are used in a complementary manner to generate pseudo-labels allowing reliable adaptation. Evaluation results reveal that S2DHand achieves significant improvements on arbitrary camera pairs under both in-dataset and cross-dataset settings and outperforms existing adaptation methods with leading performance. Project page: https://github.com/ut-vision/S2DHand.
****

Time-Efficient Light-Field Acquisition Using Coded Aperture and Events
Shuji Habuchi, Keita Takahashi, Chihiro Tsutake, Toshiaki Fujii, Hajime Nagahara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24923-24933
We propose a computational imaging method for time-efficient light-field acquisition that combines a coded aperture with an event-based camera. Different from the conventional coded-aperture imaging method our method applies a sequence of coding patterns during a single exposure for an image frame. The parallax information which is related to the differences in coding patterns is recorded as events. The image frame and events all of which are measured in a single exposure are jointly used to computationally reconstruct a light field. We also designed an algorithm pipeline for our method that is end-to-end trainable on the basis of deep optics and compatible with real camera hardware. We experimentally showed that our method can achieve more accurate reconstruction than several other imaging methods with a single exposure. We also developed a hardware prototype with the potential to complete the measurement on the camera within 22 msec and demonstrated that light fields from real 3-D scenes can be obtained with convincing visual quality. Our software and supplementary video are available from our project website.
****

EVS-assisted Joint Deblurring Rolling-Shutter Correction and Video Frame Interpolation through Sensor Inverse Modeling
Rui Jiang, Fangwen Tu, Yixuan Long, Aabhaas Vaish, Bowen Zhou, Qinyi Wang, Wei Zhang, Yuntan Fang, Luis Eduardo Garcia Capel, Bo Mu, Tiejun Dai, Andreas Suess; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25172-25181
Event-based Vision Sensors (EVS) gain popularity in enhancing CMOS Image Sensor (CIS) video capture. Nonidealities of EVS such as pixel or readout latency can significantly influence the quality of the enhanced images and warrant dedicated consideration in the design of fusion algorithms. A novel approach for jointly computing deblurred rolling-shutter artifact corrected high-speed videos with frame rates up to 10000 FPS using inherently blurry rolling shutter CIS frames of 120 FPS to 150 FPS in conjunction with EVS data from a hybrid CIS-EVS sensor is presented. EVS pixel latency readout latency and the sensor's refractory period are explicitly incorporated into the measurement model. This inverse function problem is solved on a per-pixel manner using an optimization-based framework. The interpolated images are subsequently processed by a novel refinement network. The proposed method is evaluated using simulated and measured datasets under natur

al and controlled environments. Extensive experiments show reduced shadowing eff ect a 4 dB increment in PSNR and a 12% improvement in LPIPS score compared to st ate-of-the-art methods.
****
Prompt-Enhanced Multiple Instance Learning for Weakly Supervised Video Anomaly D etection
Junxi Chen, Liang Li, Li Su, Zheng-jun Zha, Qingming Huang; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18319-18329
Weakly-supervised Video Anomaly Detection (wVAD) aims to detect frame-level anom alies using only video-level labels in training. Due to the limitation of coarse -grained labels Multi-Instance Learning (MIL) is prevailing in wVAD. However MIL suffers from insufficiency of binary supervision to model diverse abnormal patt erns. Besides the coupling between abnormality and its context hinders the learn ing of clear abnormal event boundary. In this paper we propose prompt-enhanced M IL to detect various abnormal events while ensuring clear event boundaries. Conc retely we design the abnormal-aware prompts by using abnormal class annotations together with learnable prompt which can incorporate semantic priors into video features dynamically. The detector can utilize the semantic-rich features to cap ture diverse abnormal patterns. In addition normal context prompt is introduced to amplify the distinction between abnormality and its context facilitating the generation of clear boundary. With the mutual enhancement of abnormal-aware and normal context prompt the model can construct discriminative representations to detect divergent anomalies without ambiguous event boundaries. Extensive experim ents demonstrate our method achieves SOTA performance on three public benchmarks . The code is available at https://github.com/Junxi-Chen/PE-MIL.
****
Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Charact er Animation
Li Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2024, pp. 8153-8163
Character Animation aims to generating character videos from still images throug h driving signals. Currently diffusion models have become the mainstream in visu al generation research owing to their robust generative capabilities. However ch allenges persist in the realm of image-to-video especially in character animatio n where temporally maintaining consistency with detailed information from charac ter remains a formidable problem. In this paper we leverage the power of diffusi on models and propose a novel framework tailored for character animation. To pre serve consistency of intricate appearance features from reference image we desig n ReferenceNet to merge detail features via spatial attention. To ensure control lability and continuity we introduce an efficient pose guider to direct characte r's movements and employ an effective temporal modeling approach to ensure smoot h inter-frame transitions between video frames. By expanding the training data o ur approach can animate arbitrary characters yielding superior results in charac ter animation compared to other image-to-video methods. Furthermore we evaluate our method on image animation benchmarks achieving state-of-the-art results.
****
FreeCustom: Tuning-Free Customized Image Generation for Multi-Concept Compositio n
Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2024, pp. 9089-9098
Benefiting from large-scale pre-trained text-to-image (T2I) generative models im pressive progress has been achieved in customized image generation which aims to generate user-specified concepts. Existing approaches have extensively focused on single-concept customization and still encounter challenges when it comes to complex scenarios that involve combining multiple concepts. These approaches oft en require retraining/fine-tuning using a few images leading to time-consuming t raining processes and impeding their swift implementation. Furthermore the relia nce on multiple images to represent a singular concept increases the difficulty

of customization. To this end we propose FreeCustom a novel tuning-free method to generate customized images of multi-concept composition based on reference concepts using only one image per concept as input. Specifically we introduce a new multi-reference self-attention (MRSA) mechanism and a weighted mask strategy that enables the generated image to access and focus more on the reference concepts. In addition MRSA leverages our key finding that input concepts are better preserved when providing images with context interactions. Experiments show that our method's produced images are consistent with the given concepts and better aligned with the input text. Our method outperforms or performs on par with other training-based methods in terms of multi-concept composition and single-concept customization but is simpler. Codes can be found \href https://github.com/aim-uofa/FreeCustom here .
****
Non-autoregressive Sequence-to-Sequence Vision-Language Models
Kunyu Shi, Qi Dong, Luis Goncalves, Zhuowen Tu, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13603-13612
Sequence-to-sequence vision-language models are showing promise but their applicability is limited by their inference latency due to their autoregressive way of generating predictions. We propose a parallel decoding sequence-to-sequence vision-language model trained with a Query-CTC loss that marginalizes over multiple inference paths in the decoder. This allows us to model the joint distribution of tokens rather than restricting to conditional distribution as in an autoregressive model. The resulting model NARVL achieves performance on-par with its state-of-the-art autoregressive counterpart but is faster at inference time reducing from the linear complexity associated with the sequential generation of tokens to a paradigm of constant time joint inference.
****
MaskINT: Video Editing via Interpolative Non-autoregressive Masked Transformers
Haoyu Ma, Shahin Mahdizadehaghdam, Bichen Wu, Zhipeng Fan, Yuchao Gu, Wenliang Zhao, Lior Shapira, Xiaohui Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7403-7412
Recent advances in generative AI have significantly enhanced image and video editing particularly in the context of text prompt control. State-of-the-art approaches predominantly rely on diffusion models to accomplish these tasks. However the computational demands of diffusion-based methods are substantial often necessitating large-scale paired datasets for training and therefore challenging the deployment in real applications. To address these issues this paper breaks down the text-based video editing task into two stages. First we leverage an pre-trained text-to-image diffusion model to simultaneously edit few keyframes in an zero-shot way. Second we introduce an efficient model called MaskINT which is built on non-autoregressive masked generative transformers and specializes in frame interpolation between the edited keyframes using the structural guidance from intermediate frames. Experimental results suggest that our MaskINT achieves comparable performance with diffusion-based methodologies while significantly improve the inference time. This research offers a practical solution for text-based video editing and showcases the potential of non-autoregressive masked generative transformers in this domain.
****
Active Prompt Learning in Vision Language Models
Jihwan Bang, Sumyeong Ahn, Jae-Gil Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27004-27014
Pre-trained Vision Language Models (VLMs) have demonstrated notable progress in various zero-shot tasks such as classification and retrieval. Despite their performance because improving performance on new tasks requires task-specific knowledge their adaptation is essential. While labels are needed for the adaptation acquiring them is typically expensive. To overcome this challenge active learning a method of achieving a high performance by obtaining labels for a small number of samples from experts has been studied. Active learning primarily focuses on selecting unlabeled samples for labeling and leveraging them to train models. In

this study we pose the question "how can the pre-trained VLMs be adapted under the active learning framework?" In response to this inquiry we observe that (1) simply applying a conventional active learning framework to pre-trained VLMs even may degrade performance compared to random selection because of the class imbalance in labeling candidates and (2) the knowledge of VLMs can provide hints for achieving the balance before labeling. Based on these observations we devise a novel active learning framework for VLMs denoted as PCB. To assess the effectiveness of our approach we conduct experiments on seven different real-world datasets and the results demonstrate that PCB surpasses conventional active learning and random sampling methods.
****
Learning Multi-Dimensional Human Preference for Text-to-Image Generation
Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, Zhongyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8018-8027
Current metrics for text-to-image models typically rely on statistical metrics which inadequately represent the real preference of humans. Although recent work attempts to learn these preferences via human annotated images they reduce the rich tapestry of human preference to a single overall score. However the preference results vary when humans evaluate images with different aspects. Therefore to learn the multi-dimensional human preferences we propose the Multi-dimensional Preference Score (MPS) the first multi-dimensional preference scoring model for the evaluation of text-to-image models. The MPS introduces the preference condition module upon CLIP model to learn these diverse preferences. It is trained based on our Multi-dimensional Human Preference (MHP) Dataset which comprises 918315 human preference choices across four dimensions (i.e. aesthetics semantic alignment detail quality and overall assessment) on 607541 images. The images are generated by a wide range of latest text-to-image models. The MPS outperforms existing scoring methods across 3 datasets in 4 dimensions enabling it a promising metric for evaluating and improving text-to-image generation. The model and dataset will be made publicly available to facilitate future research.
****
ViVid-1-to-3: Novel View Synthesis with Video Diffusion Models
Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, Kwang Moo Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6775-6785
Generating novel views of an object from a single image is a challenging task. It requires an understanding of the underlying 3D structure of the object from an image and rendering high-quality spatially consistent new views. While recent methods for view synthesis based on diffusion have shown great progress achieving consistency among various view estimates and at the same time abiding by the desired camera pose remains a critical problem yet to be solved. In this work we demonstrate a strikingly simple method where we utilize a pre-trained video diffusion model to solve this problem. Our key idea is that synthesizing a novel view could be reformulated as synthesizing a video of a camera going around the object of interest---a scanning video---which then allows us to leverage the powerful priors that a video diffusion model would have learned. Thus to perform novel-view synthesis we create a smooth camera trajectory to the target view that we wish to render and denoise using both a view-conditioned diffusion model and a video diffusion model. By doing so we obtain a highly consistent novel view synthesis outperforming the state of the art.
****
Active Object Detection with Knowledge Aggregation and Distillation from Large Models
Dejie Yang, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16624-16633
Accurately detecting active objects undergoing state changes is essential for comprehending human interactions and facilitating decision-making. The existing methods for active object detection (AOD) primarily rely on visual appearance of the objects within input such as changes in size shape and relationship with hand

s. However these visual changes can be subtle posing challenges particularly in scenarios with multiple distracting no-change instances of the same category. We observe that the state changes are often the result of an interaction being performed upon the object thus propose to use informed priors about object related plausible interactions (including semantics and visual appearance) to provide more reliable cues for AOD. Specifically we propose a knowledge aggregation procedure to integrate the aforementioned informed priors into oracle queries within the teacher decoder offering more object affordance commonsense to locate the active object. To streamline the inference process and reduce extra knowledge inputs we propose a knowledge distillation approach that encourages the student decoder to mimic the detection capabilities of the teacher decoder using the oracle query by replicating its predictions and attention. Our proposed framework achieves state-of-the-art performance on four datasets namely Ego4D Epic-Kitchens MECCANO and 100DOH which demonstrates the effectiveness of our approach in improving AOD. The code and models are available at https://github.com/idejie/KAD.git.
****

NICE: Neurogenesis Inspired Contextual Encoding for Replay-free Class Incremental Learning

Mustafa Burak Gurbuz, Jean Michael Moorman, Constantine Dovrolis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23659-23669

Deep neural networks (DNNs) struggle to learn in dynamic settings because they mainly rely on static datasets. Continual learning (CL) aims to overcome this limitation by enabling DNNs to incrementally accumulate knowledge. A widely adopted scenario in CL is class-incremental learning (CIL) where DNNs are required to sequentially learn more classes. Among the various strategies in CL replay methods which revisit previous classes stand out as the only effective ones in CIL. Other strategies such as architectural modifications to segregate information across weights and protect them from change are ineffective in CIL. This is because they need additional information during testing to select the correct network parts to use. In this paper we propose NICE Neurogenesis Inspired Contextual Encoding a replay-free architectural method inspired by adult neurogenesis in the hippocampus. NICE groups neurons in the DNN based on different maturation stages and infers which neurons to use during testing without any additional signal. Through extensive experiments across 6 datasets and 3 architectures we show that NICE performs on par with or often outperforms replay methods. We also make the case that neurons exhibit highly distinctive activation patterns for the classes in which they specialize enabling us to determine when they should be used. The code is available at https://github.com/BurakGurbuz97/NICE.
****

Generating Human Motion in 3D Scenes from Text Descriptions

Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1855-1866

Generating human motions from textual descriptions has gained growing research interest due to its wide range of applications. However only a few works consider human-scene interactions together with text conditions which is crucial for visual and physical realism. This paper focuses on the task of generating human motions in 3D indoor scenes given text descriptions of the human-scene interactions. This task presents challenges due to the multimodality nature of text scene and motion as well as the need for spatial reasoning. To address these challenges we propose a new approach that decomposes the complex problem into two more manageable sub-problems: (1) language grounding of the target object and (2) object-centric motion generation. For language grounding of the target object we leverage the power of large language models. For motion generation we design an object-centric scene representation for the generative model to focus on the target object thereby reducing the scene complexity and facilitating the modeling of the relationship between human motions and the object. Experiments demonstrate the better motion quality of our approach compared to baselines and validate our design choices. Code will be available at https://zju3dv.github.io/text_scene_motion

.
****
Weak-to-Strong 3D Object Detection with X-Ray Distillation
Alexander Gambashidze, Aleksandr Dadukin, Maxim Golyadkin, Maria Razzhivina, Ilya Makarov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15055-15064
This paper addresses the critical challenges of sparsity and occlusion in LiDAR-based 3D object detection. Current methods often rely on supplementary modules or specific architectural designs potentially limiting their applicability to new and evolving architectures. To our knowledge we are the first to propose a versatile technique that seamlessly integrates into any existing framework for 3D Object Detection marking the first instance of Weak-to-Strong generalization in 3D computer vision. We introduce a novel framework X-Ray Distillation with Object-Complete Frames suitable for both supervised and semi-supervised settings that leverages the temporal aspect of point cloud sequences. This method extracts crucial information from both previous and subsequent LiDAR frames creating Object-Complete frames that represent objects from multiple viewpoints thus addressing occlusion and sparsity. Given the limitation of not being able to generate Object-Complete frames during online inference we utilize Knowledge Distillation within a Teacher-Student framework. This technique encourages the strong Student model to emulate the behavior of the weaker Teacher which processes simple and informative Object-Complete frames effectively offering a comprehensive view of objects as if seen through X-ray vision. Our proposed methods surpass state-of-the-art in semi-supervised learning by 1-1.5 mAP and enhance the performance of five established supervised models by 1-2 mAP on standard autonomous driving datasets even with default hyperparameters. Code for Object-Complete frames is available here: https://github.com/sakharok13/X-Ray-Teacher-Patching-Tools.
****
QDFormer: Towards Robust Audiovisual Segmentation in Complex Environments with Quantization-based Semantic Decomposition
Xiang Li, Jinglu Wang, Xiaohao Xu, Xiulian Peng, Rita Singh, Yan Lu, Bhiksha Raj; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3402-3413
Audiovisual segmentation (AVS) is a challenging task that aims to segment visual objects in videos according to their associated acoustic cues. With multiple sound sources and background disturbances involved establishing robust correspondences between audio and visual contents poses unique challenges due to (1) complex entanglement across sound sources and (2) frequent changes in the occurrence of distinct sound events. Assuming sound events occur independently the multi-source semantic space can be represented as the Cartesian product of single-source sub-spaces. We are motivated to decompose the multi-source audio semantics into single-source semantics for more effective interactions with visual content. We propose a semantic decomposition method based on product quantization where the multi-source semantics can be decomposed and represented by several disentangled and noise-suppressed single-source semantics. Furthermore we introduce a global-to-local quantization mechanism which distills knowledge from stable global (clip-level) features into local (frame-level) ones to handle frequent changes in audio semantics. Extensive experiments demonstrate that our semantically decomposed audio representation significantly improves AVS performance eg +21.2% mIoU on the challenging AVS-Semantic benchmark with ResNet50 backbone.
****
Active Open-Vocabulary Recognition: Let Intelligent Moving Mitigate CLIP Limitations
Lei Fan, Jianxiong Zhou, Xiaoying Xing, Ying Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16394-16403
Active recognition which allows intelligent agents to explore observations for better recognition performance serves as a prerequisite for various embodied AI tasks such as grasping navigation and room arrangements. Given the evolving environment and the multitude of object classes it is impractical to include all possible classes during the training stage. In this paper we aim at advancing active

open-vocabulary recognition empowering embodied agents to actively perceive and
classify arbitrary objects. However directly adopting recent open-vocabulary cl
assification models like Contrastive Language Image Pretraining (CLIP) poses its
unique challenges. Specifically we observe that CLIP's performance is heavily a
ffected by the viewpoint and occlusions compromising its reliability in unconstr
ained embodied perception scenarios. Further the sequential nature of observatio
ns in agent-environment interactions necessitates an effective method for integr
ating features that maintains discriminative strength for open-vocabulary classi
fication. To address these issues we introduce a novel agent for active open-voc
abulary recognition. The proposed method leverages inter-frame and inter-concept
similarities to navigate agent movements and to fuse features without relying o
n class-specific knowledge. Compared to baseline CLIP model with 29.6% accuracy
on ShapeNet dataset the proposed agent could achieve 53.3% accuracy for open-voc
abulary recognition without any fine-tuning to the equipped CLIP model. Addition
al experiments conducted with the Habitat simulator further affirm the efficacy
of our method.
****
Backdoor Defense via Test-Time Detecting and Repairing
Jiyang Guan, Jian Liang, Ran He; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2024, pp. 24564-24573
Deep neural networks have played a crucial part in many critical domains such as
autonomous driving face recognition and medical diagnosis. However deep neural
networks are facing security threats from backdoor attacks and can be manipulate
d into attacker-decided behaviors by the backdoor attacker. To defend the backdo
or prior research has focused on using clean data to remove backdoor attacks bef
ore model deployment. In this paper we investigate the possibility of defending
against backdoor attacks by utilizing test-time partially poisoned data to remov
e the backdoor from the model. To address the problem a two-stage method TTBD is
proposed. In the first stage we propose a backdoor sample detection method DDP
to identify poisoned samples from a batch of mixed partially poisoned samples. O
nce the poisoned samples are detected we employ Shapley estimation to calculate
the contribution of each neuron's significance in the network locate the poisone
d neurons and prune them to remove backdoor in the models. Our experiments demon
strate that TTBD removes the backdoor successfully with only a batch of partiall
y poisoned data across different model architectures and datasets against differ
ent types of backdoor attacks.
****
Fast Adaptation for Human Pose Estimation via Meta-Optimization
Shengxiang Hu, Huaijiang Sun, Bin Li, Dong Wei, Weiqing Li, Jianfeng Lu; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 1792-1801
Domain shift is a challenge for supervised human pose estimation where the sourc
e data and target data come from different distributions. This is why pose estim
ation methods generally perform worse on the test set than on the training set.
Recently test-time adaptation has proven to be an effective way to deal with dom
ain shift in human pose estimation. Although the performance on the target domai
n has been improved existing methods require a large number of weight updates fo
r convergence which is time-consuming and brings catastrophic forgetting. To sol
ve these issues we propose a meta-auxiliary learning method to achieve fast adap
tation for domain shift during inference. Specifically we take human pose estima
tion as the supervised primary task and propose body-specific image inpainting a
s a self-supervised auxiliary task. First we jointly train the primary and auxil
iary tasks to get a pre-trained model on the source domain. Then meta-training c
orrelates the performance of the two tasks to learn a good weight initialization
. Finally meta-testing adapts the meta-learned model to the target data through
self-supervised learning. Benefiting from the meta-learning paradigm the propose
d method enables fast adaptation to the target domain while preserving the sourc
e domain knowledge. The carefully designed auxiliary task better pays attention
to human-related semantics in a single image. Extensive experiments demonstrate
the effectiveness of our test-time fast adaptation.

****
Efficient Meshflow and Optical Flow Estimation from Event Cameras
Xinglong Luo, Ao Luo, Zhengning Wang, Chunyu Lin, Bing Zeng, Shuaicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19198-19207
In this paper we explore the problem of event-based meshflow estimation a novel task that involves predicting a spatially smooth sparse motion field from event cameras. To start we generate a large-scale High-Resolution Event Meshflow (HREM) dataset which showcases its superiority by encompassing the merits of high resolution at 1280x720 handling dynamic objects and complex motion patterns and offering both optical flow and meshflow labels. These aspects have not been fully explored in previous works. Besides we propose Efficient Event-based MeshFlow (EEMFlow) network a lightweight model featuring a specially crafted encoder-decoder architecture to facilitate swift and accurate meshflow estimation. Furthermore we upgrade EEMFlow network to support dense event optical flow in which a Confidence-induced Detail Completion (CDC) module is proposed to preserve sharp motion boundaries. We conduct comprehensive experiments to show the exceptional performance and runtime efficiency (39x faster) of our EEMFlow model compared to recent state-of-the-art flow methods. Our code is available at https://github.com/boomluo02/EEMFlow.
****
Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models
Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, Ariel Fuxman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9590-9601
Solving complex visual tasks such as "Who invented the musical instrument on the right?" involves a composition of skills: understanding space recognizing instruments and also retrieving prior knowledge. Recent work shows promise by decomposing such tasks using a large language model (LLM) into an executable program that invokes specialized vision models. However generated programs are error-prone: they omit necessary steps include spurious ones and are unable to recover when the specialized models give incorrect outputs. Moreover they require loading multiple models incurring high latency and computation costs. We propose Visual Program Distillation (VPD) an instruction-tuning framework that produces a vision-language model (VLM) capable of solving complex visual tasks with a single forward pass. VPD distills the reasoning ability of LLMs by using them to sample multiple candidate programs which are then executed and verified to identify the correct one. It translates each correct program into a language description of the reasoning steps which are then distilled into a VLM. Extensive experiments show that VPD improves the VLM's ability to count understand spatial relations and reason compositionally. Our VPD-trained PaLI-X outperforms all prior VLMs achieving state-of-the-art performance across complex vision tasks including MMBench OK-VQA A-OKVQA TallyQA POPE and Hateful Memes. An evaluation with human annotators also confirms that VPD improves model response factuality and consistency. Finally experiments on content moderation demonstrate that VPD is also helpful for adaptation to real-world applications with limited data.
****
OneFormer3D: One Transformer for Unified Point Cloud Segmentation
Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, Danila Rukhovich; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20943-20953
Semantic instance and panoptic segmentation of 3D point clouds have been addressed using task-specific models of distinct design. Thereby the similarity of all segmentation tasks and the implicit relationship between them have not been utilized effectively. This paper presents a unified simple and effective model addressing all these tasks jointly. The model named OneFormer3D performs instance and semantic segmentation consistently using a group of learnable kernels where each kernel is responsible for generating a mask for either an instance or a semantic category. These kernels are trained with a transformer-based decoder with uni

fied instance and semantic queries passed as an input. Such a design enables tra
ining a model end-to-end in a single run so that it achieves top performance on
all three segmentation tasks simultaneously. Specifically our OneFormer3D ranks
1st and sets a new state-of-the-art (+2.1 mAP50) in the ScanNet test leaderboard
. We also demonstrate the state-of-the-art results in semantic instance and pano
ptic segmentation of ScanNet (+21 PQ) ScanNet200 (+3.8 mAP50) and S3DIS (+0.8 mI
oU) datasets.
****
JRDB-Social: A Multifaceted Robotic Dataset for Understanding of Context and Dyn
amics of Human Interactions Within Social Groups
Simindokht Jahangard, Zhixi Cai, Shiki Wen, Hamid Rezatofighi; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 22087-22097
Understanding human social behaviour is crucial in computer vision and robotics.
 Micro-level observations like individual actions fall short necessitating a com
prehensive approach that considers individual behaviour intra-group dynamics and
 social group levels for a thorough understanding. To address dataset limitation
s this paper introduces JRDB-Social an extension of JRDB. Designed to fill gaps
in human understanding across diverse indoor and outdoor social contexts JRDB-So
cial provides annotations at three levels: individual attributes intra-group int
eractions and social group context. This dataset aims to enhance our grasp of hu
man social dynamics for robotic applications. Utilizing the recent cutting-edge
multi-modal large language models we evaluated our benchmark to explore their ca
pacity to decipher social human behaviour.
****
A Backpack Full of Skills: Egocentric Video Understanding with Diverse Task Pers
pectives
Simone Alberto Peirone, Francesca Pistilli, Antonio Alliegro, Giuseppe Averta; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 18275-18285
Human comprehension of a video stream is naturally broad: in a few instants we a
re able to understand what is happening the relevance and relationship of object
s and forecast what will follow in the near future everything all at once. We be
lieve that - to effectively transfer such an holistic perception to intelligent
machines - an important role is played by learning to correlate concepts and to
abstract knowledge coming from different tasks to synergistically exploit them w
hen learning novel skills. To accomplish this we look for a unified approach to
video understanding which combines shared temporal modelling of human actions wi
th minimal overhead to support multiple downstream tasks and enable cooperation
when learning novel skills. We then propose EgoPack a solution that creates a co
llection of task perspectives that can be carried across downstream tasks and us
ed as a potential source of additional insights as a backpack of skills that a r
obot can carry around and use when needed. We demonstrate the effectiveness and
efficiency of our approach on four Ego4D benchmarks outperforming current state-
of-the-art methods. Project webpage: https://sapeirone.github.io/EgoPack.
****
WOUAF: Weight Modulation for User Attribution and Fingerprinting in Text-to-Imag
e Diffusion Models
Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, Yezhou Yang; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 8974-8983
The rapid advancement of generative models facilitating the creation of hyper-re
alistic images from textual descriptions has concurrently escalated critical soc
ietal concerns such as misinformation. Although providing some mitigation tradit
ional fingerprinting mechanisms fall short in attributing responsibility for the
 malicious use of synthetic images. This paper introduces a novel approach to mo
del fingerprinting that assigns responsibility for the generated images thereby
serving as a potential countermeasure to model misuse. Our method modifies gener
ative models based on each user's unique digital fingerprint imprinting a unique
 identifier onto the resultant content that can be traced back to the user. This

approach incorporating fine-tuning into Text-to-Image (T2I) tasks using the Sta
ble Diffusion Model demonstrates near-perfect attribution accuracy with a minima
l impact on output quality. Through extensive evaluation we show that our method
 outperforms baseline methods with an average improvement of 11% in handling ima
ge post-processes. Our method presents a promising and novel avenue for accounta
ble model distribution and responsible use. Our code is available in https://git
hub.com/kylemin/WOUAF.
****
Visual In-Context Prompting
Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu,
 Hongyang Li, Jianwei Yang, Chunyuan Li, Lei Zhang, Jianfeng Gao; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 12861-12871
In-context prompting in large language models (LLMs) has become a prevalent appr
oach to improve zero-shot capabilities but this idea is less explored in the vis
ion domain. Existing visual prompting methods focus on referring segmentation to
 segment the most relevant object falling short of addressing many generic visio
n tasks like open-set segmentation and detection. In this paper we introduce a u
niversal visual in-context prompting framework for both tasks as shown in Fig.1.
 In particular we build on top of an encoder-decoder architecture and develop a
versatile prompt encoder to support a variety of prompts like strokes boxes and
points. We further enhance it to take an arbitrary number of reference image seg
ments as the context. Our extensive explorations show that the proposed visual i
n-context prompting elicits extraordinary referring and generic segmentation cap
abilities to refer and detect yielding competitive performance to close-set in-d
omain datasets and showing promising results on many open-set segmentation datas
ets. By joint training on COCO and SA-1B DINOv achieves 57.7 PQ on COCO and 23.2
 PQ on ADE20K. Code will be available at https://github.com/UX-Decoder/DINOv
****
Text-Conditioned Generative Model of 3D Strand-based Human Hairstyles
Vanessa Sklyarova, Egor Zakharov, Otmar Hilliges, Michael J. Black, Justus Thies
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2024, pp. 4703-4712
We present HAAR a new strand-based generative model for 3D human hairstyles. Spe
cifically based on textual inputs HAAR produces 3D hairstyles that could be used
 as production-level assets in modern computer graphics engines. Current AI-base
d generative models take advantage of powerful 2D priors to reconstruct 3D conte
nt in the form of point clouds meshes or volumetric functions. However by using
the 2D priors they are intrinsically limited to only recovering the visual parts
. Highly occluded hair structures can not be reconstructed with those methods an
d they only model the "outer shell" which is not ready to be used in physics-bas
ed rendering or simulation pipelines. In contrast we propose a first text-guided
 generative method that uses 3D hair strands as an underlying representation. Le
veraging 2D visual question-answering (VQA) systems we automatically annotate sy
nthetic hair models that are generated from a small set of artist-created hairst
yles. This allows us to train a latent diffusion model that operates in a common
 hairstyle UV space. In qualitative and quantitative studies we demonstrate the
capabilities of the proposed model and compare it to existing hairstyle generati
on approaches. For results please refer to our project page https://haar.is.tue.
mpg.de/.
****
GPT-4V(ision) is a Human-Aligned Evaluator for Text-to-3D Generation
Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua
Lin, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2024, pp. 22227-22238
Despite recent advances in text-to-3D generative methods there is a notable abse
nce of reliable evaluation metrics. Existing metrics usually focus on a single c
riterion each such as how well the asset aligned with the input text. These metr
ics lack the flexibility to generalize to different evaluation criteria and migh
t not align well with human preferences. Conducting user preference studies is a

n alternative that offers both adaptability and human-aligned results. User studies however can be very expensive to scale. This paper presents an automatic versatile and human-aligned evaluation metric for text-to-3D generative models. To this end we first develop a prompt generator using GPT-4V to generate evaluating prompts which serve as input to compare text-to-3D models. We further design a method instructing GPT-4V to compare two 3D assets according to user-defined criteria. Finally we use these pairwise comparison results to assign these models Elo ratings. Experimental results suggest our metric strongly align with human preference across different evaluation criteria.
****
NTO3D: Neural Target Object 3D Reconstruction with Segment Anything
Xiaobao Wei, Renrui Zhang, Jiarui Wu, Jiaming Liu, Ming Lu, Yandong Guo, Shanghang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20352-20362
Neural 3D reconstruction from multi-view images has recently attracted increasing attention from the community. Existing methods normally learn a neural field for the whole scene while it is still under-explored how to reconstruct a target object indicated by users. Considering the Segment Anything Model (SAM) has shown effectiveness in segmenting any 2D images in this paper we propose NTO3D a novel high-quality Neural Target Object 3D (NTO3D) reconstruction method which leverages the benefits of both neural field and SAM. We first propose a novel strategy to lift the multi-view 2D segmentation masks of SAM into a unified 3D occupancy field. The 3D occupancy field is then projected into 2D space and generates the new prompts for SAM. This process is iterative until convergence to separate the target object from the scene. After this we then lift the 2D features of the SAM encoder into a 3D feature field in order to improve the reconstruction quality of the target object. NTO3D lifts the 2D masks and features of SAM into the 3D neural field for high-quality neural target object 3D reconstruction. We conduct detailed experiments on several benchmark datasets to demonstrate the advantages of our method. The code will be available at: https://github.com/ucwxb/NTO3D.
****
Instruct-ReID: A Multi-purpose Person Re-identification Task with Instructions
Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, Donglian Qi, Yunfeng Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17521-17531
Human intelligence can retrieve any person according to both visual and language descriptions. However the current computer vision community studies specific person re-identification (ReID) tasks in different scenarios separately which limits the applications in the real world. This paper strives to resolve this problem by proposing a new instruct-ReID task that requires the model to retrieve images according to the given image or language instructions. Our instruct-ReID is a more general ReID setting where existing 6 ReID tasks can be viewed as special cases by designing different instructions. We propose a large-scale OmniReID benchmark and an adaptive triplet loss as a baseline method to facilitate research in this new setting. Experimental results show that the proposed multi-purpose ReID model trained on our OmniReID benchmark without finetuning can improve +0.5% +0.6% +7.7% mAP on Market1501 MSMT17 CUHK03 for traditional ReID +6.4% +7.1% +11.2% mAP on PRCC VC-Clothes LTCC for clothes-changing ReID +11.7% mAP on COCAS+ real2 for clothes template based clothes-changing ReID when using only RGB images +24.9% mAP on COCAS+ real2 for our newly defined language-instructed ReID +4.3% on LLCM for visible-infrared ReID +2.6% on CUHK-PEDES for text-to-image ReID. The datasets the model and code are available at https://github.com/hwz-zju/Instruct-ReID.
****
OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM
Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (

CVPR), 2024, pp. 22170-22183
Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities i
n various multimodal tasks. However their potential in the medical domain remain
s largely unexplored. A significant challenge arises from the scarcity of divers
e medical images spanning various modalities and anatomical regions which is ess
ential in real-world medical applications. To solve this problem in this paper w
e introduce OmniMedVQA a novel comprehensive medical Visual Question Answering (
VQA) benchmark. This benchmark is collected from 73 different medical datasets i
ncluding 12 different modalities and covering more than 20 distinct anatomical r
egions. Importantly all images in this benchmark are sourced from authentic medi
cal scenarios ensuring alignment with the requirements of the medical field and
suitability for evaluating LVLMs. Through our extensive experiments we have foun
d that existing LVLMs struggle to address these medical VQA problems effectively
. Moreover what surprises us is that medical-specialized LVLMs even exhibit infe
rior performance to those general-domain models calling for a more versatile and
 robust LVLM in the biomedical field. The evaluation results not only reveal the
 current limitations of LVLM in understanding real medical images but also highl
ight our dataset's significance. Our code with dataset are available at https://
github.com/OpenGVLab/Multi-Modality-Arena.
****
Skeleton-in-Context: Unified Skeleton Sequence Modeling with In-Context Learning
Xinshun Wang, Zhongbin Fang, Xia Li, Xiangtai Li, Chen Chen, Mengyuan Liu; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 2436-2446
In-context learning provides a new perspective for multi-task modeling for visio
n and NLP. Under this setting the model can perceive tasks from prompts and acco
mplish them without any extra task-specific head predictions or model fine-tunin
g. However skeleton sequence modeling via in-context learning remains unexplored
. Directly applying existing in-context models from other areas onto skeleton se
quences fails due to the similarity between inter-frame and cross-task poses whi
ch makes it exceptionally hard to perceive the task correctly from a subtle cont
ext. To address this challenge we propose Skeleton-in-Context (SiC) an effective
 framework for in-context skeleton sequence modeling. Our SiC is able to handle
multiple skeleton-based tasks simultaneously after a single training process and
 accomplish each task from context according to the given prompt. It can further
 generalize to new unseen tasks according to customized prompts. To facilitate c
ontext perception we additionally propose a task-unified prompt which adaptively
 learns tasks of different natures such as partial joint-level generation sequen
ce-level prediction or 2D-to-3D motion prediction. We conduct extensive experime
nts to evaluate the effectiveness of our SiC on multiple tasks including motion
prediction pose estimation joint completion and future pose estimation. We also
evaluate its generalization capability on unseen tasks such as motion-in-between
. These experiments show that our model achieves state-of-the-art multi-task per
formance and even outperforms single-task methods on certain tasks.
****
DemoFusion: Democratising High-Resolution Image Generation With No $$$
Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, Zhanyu Ma; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2024, pp. 6159-6168
High-resolution image generation with Generative Artificial Intelligence (GenAI)
 has immense potential but due to the enormous capital investment required for t
raining it is increasingly centralised to a few large corporations and hidden be
hind paywalls. This paper aims to democratise high-resolution GenAI by advancing
 the frontier of high-resolution generation while remaining accessible to a broa
d audience. We demonstrate that existing Latent Diffusion Models (LDMs) possess
untapped potential for higher-resolution image generation. Our novel DemoFusion
framework seamlessly extends open-source GenAI models employing Progressive Upsc
aling Skip Residual and Dilated Sampling mechanisms to achieve higher-resolution
 image generation. The progressive nature of DemoFusion requires more passes but
 the intermediate results can serve as "previews" facilitating rapid prompt iter

ation.
****
IBD-SLAM: Learning Image-Based Depth Fusion for Generalizable SLAM
Minghao Yin, Shangzhe Wu, Kai Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10563-10573
In this paper we address the challenging problem of visual SLAM with neural scene representations. Recently neural scene representations have shown promise for SLAM to produce dense 3D scene reconstruction with high quality. However existing methods require scene-specific optimization leading to time-consuming mapping processes for each individual scene. To overcome this limitation we propose IBD-SLAM an Image-Based Depth fusion framework for generalizable SLAM. In particular we adopt a Neural Radiance Field (NeRF) for scene representation. Inspired by multi-view image-based rendering instead of learning a fixed-grid scene representation we propose to learn an image-based depth fusion model that fuses depth maps of multiple reference views into a xyz-map representation. Once trained this model can be applied to new uncalibrated monocular RGBD videos of unseen scenes without the need for retraining and reconstructs full 3D scenes efficiently with a light-weight pose optimization procedure. We thoroughly evaluate IBD-SLAM on public visual SLAM benchmarks outperforming the previous state-of-the-art while being 10x faster in the mapping stage. Project page: https://visual-ai.github.io/ibd-slam.
****
CPLIP: Zero-Shot Learning for Histopathology with Comprehensive Vision-Language Alignment
Sajid Javed, Arif Mahmood, Iyyakutti Iyappan Ganapathi, Fayaz Ali Dharejo, Naoufel Werghi, Mohammed Bennamoun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11450-11459
This paper proposes Comprehensive Pathology Language Image Pre-training (CPLIP) a new unsupervised technique designed to enhance the alignment of images and text in histopathology for tasks such as classification and segmentation. This methodology enriches vision language models by leveraging extensive data without needing ground truth annotations. CPLIP involves constructing a pathology-specific dictionary generating textual descriptions for images using language models and retrieving relevant images for each text snippet via a pre-trained model. The model is then fine-tuned using a many-to-many contrastive learning method to align complex interrelated concepts across both modalities. Evaluated across multiple histopathology tasks CPLIP shows notable improvements in zero-shot learning scenarios outperforming existing methods in both interpretability and robustness and setting a higher benchmark for the application of vision-language models in the field. To encourage further research and replication the code for CPLIP is available on GitHubat https://cplip.github.io/
****
Total Selfie: Generating Full-Body Selfies
Bowei Chen, Brian Curless, Ira Kemelmacher-Shlizerman, Steven M. Seitz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6701-6711
We present a method to generate full-body selfies from photographs originally taken at arms length. Because self-captured photos are typically taken close up they have limited field of view and exaggerated perspective that distorts facial shapes. We instead seek to generate the photo some one else would take of you from a few feet away. Our approach takes as input four selfies of your face and body a background image and generates a full-body selfie in a desired target pose. We introduce a novel diffusion-based approach to combine all of this information into high-quality well-composed photos of you with the desired pose and background.
****
Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding
Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, Zhen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20623-20633

3D Visual Grounding (3DVG) aims at localizing 3D object based on textual descrip tions. Conventional supervised methods for 3DVG often necessitate extensive anno tations and a predefined vocabulary which can be restrictive. To address this is sue we propose a novel visual programming approach for zero-shot open-vocabulary 3DVG leveraging the capabilities of large language models (LLMs). Our approach begins with a unique dialog-based method engaging with LLMs to establish a found ational understanding of zero-shot 3DVG. Building on this we design a visual pro gram that consists of three types of modules i.e. view-independent view-dependen t and functional modules. Furthermore we develop an innovative language-object c orrelation module to extend the scope of existing 3D object detectors into open-vocabulary scenarios. Extensive experiments demonstrate that our zero-shot appro ach can outperform some supervised baselines marking a significant stride toward s effective 3DVG. Code is available at https://curryyuan.github.io/ZSVG3D.
****
Learning Structure-from-Motion with Graph Attention Networks
Lucas Brynte, José Pedro Iglesias, Carl Olsson, Fredrik Kahl; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4808-4817
In this paper we tackle the problem of learning Structure-from-Motion (SfM) thro ugh the use of graph attention networks. SfM is a classic computer vision proble m that is solved though iterative minimization of reprojection errors referred t o as Bundle Adjustment (BA) starting from a good initialization. In order to obt ain a good enough initialization to BA conventional methods rely on a sequence o f sub-problems (such as pairwise pose estimation pose averaging or triangulation) which provide an initial solution that can then be refined using BA. In this w ork we replace these sub-problems by learning a model that takes as input the 2D keypoints detected across multiple views and outputs the corresponding camera p oses and 3D keypoint coordinates. Our model takes advantage of graph neural netw orks to learn SfM-specific primitives and we show that it can be used for fast i nference of the reconstruction for new and unseen sequences. The experimental re sults show that the proposed model outperforms competing learning-based methods and challenges COLMAP while having lower runtime. Our code is available at: http s://github.com/lucasbrynte/gasfm/.
****
Geometry Transfer for Stylizing Radiance Fields
Hyunyoung Jung, Seonghyeon Nam, Nikolaos Sarafianos, Sungjoo Yoo, Alexander Sork ine-Hornung, Rakesh Ranjan; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2024, pp. 8565-8575
Shape and geometric patterns are essential in defining stylistic identity. Howev er current 3D style transfer methods predominantly focus on transferring colors and textures often overlooking geometric aspects. In this paper we introduce Geo metry Transfer a novel method that leverages geometric deformation for 3D style transfer. This technique employs depth maps to extract a style guide subsequentl y applied to stylize the geometry of radiance fields. Moreover we propose new te chniques that utilize geometric cues from the 3D scene thereby enhancing aesthet ic expressiveness and more accurately reflecting intended styles. Our extensive experiments show that Geometry Transfer enables a broader and more expressive ra nge of stylizations thereby significantly expanding the scope of 3D style transf er.
****
Holoported Characters: Real-time Free-viewpoint Rendering of Humans from Sparse RGB Cameras
Ashwath Shetty, Marc Habermann, Guoxing Sun, Diogo Luvizon, Vladislav Golyanik, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2024, pp. 1206-1215
We present the first approach to render highly realistic free-viewpoint videos o f a human actor in general apparel from sparse multi-view recording to display i n real-time at an unprecedented 4K resolution. At inference our method only requ ires four camera views of the moving actor and the respective 3D skeletal pose. It handles actors in wide clothing and reproduces even fine-scale dynamic detail

e.g. clothing wrinkles face expressions and hand gestures. At training time our
 learning-based approach expects dense multi-view video and a rigged static surf
ace scan of the actor. Our method comprises three main stages. Stage 1 is a skel
eton-driven neural approach for high-quality capture of the detailed dynamic mes
h geometry. Stage 2 is a novel solution to create a view-dependent texture using
 four test-time camera views as input. Finally stage 3 comprises a new image-bas
ed refinement network rendering the final 4K image given the output from the pre
vious stages. Our approach establishes a new benchmark for real-time rendering r
esolution and quality using sparse input camera views unlocking possibilities fo
r immersive telepresence.
****
SEAS: ShapE-Aligned Supervision for Person Re-Identification
Haidong Zhu, Pranav Budhwant, Zhaoheng Zheng, Ram Nevatia; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1
64-174
We introduce SEAS using ShapE-Aligned Supervision to enhance appearance-based pe
rson re-identification. When recognizing an individual's identity existing metho
ds primarily rely on appearance which can be influenced by the background enviro
nment due to a lack of body shape awareness. Although some methods attempt to in
corporate other modalities such as gait or body shape they encode the additional
 modality separately resulting in extra computational costs and lacking an inher
ent connection with appearance. In this paper we explore the use of implicit 3-D
 body shape representations as pixel-level guidance to augment the extraction of
 identity features with body shape knowledge in addition to appearance. Using bo
dy shape as supervision rather than as input provides shape-aware enhancements w
ithout any increase in computational cost and delivers coherent integration with
 pixel-wise appearance features. Moreover for video-based person re-identificati
on we align pixel-level features across frames with shape awareness to ensure te
mporal consistency. Our results demonstrate that incorporating body shape as pix
el-level supervision reduces rank-1 errors by 1.4% for frame-based and by 2.5% f
or video-based re-identification tasks respectively and can also be generalized
to other existing appearance-based person re-identification methods.
****
Class Incremental Learning with Multi-Teacher Distillation
Haitao Wen, Lili Pan, Yu Dai, Heqian Qiu, Lanxiao Wang, Qingbo Wu, Hongliang Li;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2024, pp. 28443-28452
Distillation strategies are currently the primary approaches for mitigating forg
etting in class incremental learning (CIL). Existing methods generally inherit p
revious knowledge from a single teacher. However teachers with different mechani
sms are talented at different tasks and inheriting diverse knowledge from them c
an enhance compatibility with new knowledge. In this paper we propose the MTD me
thod to find multiple diverse teachers for CIL. Specifically we adopt weight per
mutation feature perturbation and diversity regularization techniques to ensure
diverse mechanisms in teachers. To reduce time and memory consumption each teach
er is represented as a small branch in the model. We adapt existing CIL distilla
tion strategies with MTD and extensive experiments on CIFAR-100 ImageNet-100 and
 ImageNet-1000 show significant performance improvement. Our code is available a
t https://github.com/HaitaoWen/CLearning.
****
Reg-PTQ: Regression-specialized Post-training Quantization for Fully Quantized O
bject Detector
Yifu Ding, Weilun Feng, Chuyan Chen, Jinyang Guo, Xianglong Liu; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 16174-16184
Although deep learning based object detection is of great significance for vario
us applications it faces challenges when deployed on edge devices due to the com
putation and energy limitations. Post-training quantization (PTQ) can improve in
ference efficiency through integer computing. However they suffer from severe pe
rformance degradation when performing full quantization due to overlooking the u

nique characteristics of regression tasks in object detection. In this paper we are the first to explore regression-friendly quantization and conduct full quant ization on various detectors. We reveal the intrinsic reason behind the difficul ty of quantizing regressors with empirical and theoretical justifications and in troduce a novel Regression-specialized Post-Training Quantization (Reg-PTQ) sche me. It includes Filtered Global Loss Integration Calibration to combine the glob al loss with a two-step filtering mechanism mitigating the adverse impact of fal se positive bounding boxes and Learnable Logarithmic-Affine Quantizer tailored f or the non-uniform distributed parameters in regression structures. Extensive ex periments on prevalent detectors showcase the effectiveness of the well-designed Reg-PTQ. Notably our Reg-PTQ achieves 7.6 times and 5.4 times reduction in comp utation and storage consumption under INT4 with little performance degradation w hich indicates the immense potential of fully quantized detectors in real-world object detection applications.
****
AMU-Tuning: Effective Logit Bias for CLIP-based Few-shot Learning
Yuwei Tang, Zhenyi Lin, Qilong Wang, Pengfei Zhu, Qinghua Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp . 23323-23333
Recently pre-trained vision-language models (e.g. CLIP) have shown great potenti al in few-shot learning and attracted a lot of research interest. Although effor ts have been made to improve few-shot ability of CLIP key factors on the effecti veness of existing methods have not been well studied limiting further explorati on of CLIP's potential in few-shot learning. In this paper we first introduce a unified formulation to analyze CLIP-based few-shot learning methods from a persp ective of logit bias which encourages us to learn an effective logit bias for fu rther improving performance of CLIP-based few-shot learning methods. To this end we disassemble three key components involved in computation of logit bias (i.e. logit features logit predictor and logit fusion) and empirically analyze the ef fect on performance of few-shot classification. Based on analysis of key compone nts this paper proposes a novel AMU-Tuning method to learn effective logit bias for CLIP-based few-shot classification. Specifically our AMU-Tuning predicts log it bias by exploiting the appropriate Auxiliary features which are fed into an e fficient feature-initialized linear classifier with Multi-branch training. Final ly an Uncertainty-based fusion is developed to incorporate logit bias into CLIP for few-shot classification. The experiments are conducted on several widely use d benchmarks and the results show AMU-Tuning clearly outperforms its counterpart s while achieving state-of-the-art performance of CLIP-based few-shot learning w ithout bells and whistles.
****
Real-World Mobile Image Denoising Dataset with Efficient Baselines
Roman Flepp, Andrey Ignatov, Radu Timofte, Luc Van Gool; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 223 68-22377
The recently increased role of mobile photography has raised the standards of on -device photo processing tremendously. Despite the latest advancements in camera hardware the mobile camera sensor area cannot be increased significantly due to physical constraints leading to a pixel size of 0.6--2.0 \mum which results in strong image noise even in moderate lighting conditions. In the era of deep lear ning one can train a CNN model to perform robust image denoising. However there is still a lack of a substantially diverse dataset for this task. To address thi s problem we introduce a novel Mobile Image Denoising Dataset (MIDD) comprising over 400000 noisy / noise-free image pairs captured under various conditions by 20 different mobile camera sensors. Additionally we propose a new DPreview test set consisting of data from 294 different cameras for precise model evaluation. Furthermore we present the efficient baseline model SplitterNet for the consider ed mobile image denoising task that achieves high numerical and visual results w hile being able to process 8MP photos directly on smartphone GPUs in under one s econd. Thereby outperforming models with similar runtimes. This model is also co mpatible with recent mobile NPUs demonstrating an even higher speed when deploye

d on them. The conducted experiments demonstrate high robustness of the proposed solution when applied to images from previously unseen sensors showing its high generalizability. The datasets code and models can be found on the official project website.
****
Making Vision Transformers Truly Shift-Equivariant
Renan A. Rojas-Gomez, Teck-Yian Lim, Minh N. Do, Raymond A. Yeh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5568-5577
In the field of computer vision Vision Transformers (ViTs) have emerged as a prominent deep learning architecture. Despite being inspired by Convolutional Neural Networks (CNNs) ViTs are susceptible to small spatial shifts in the input data - they lack shift-equivariance. To address this shortcoming we introduce novel data-adaptive designs for each of the ViT modules that break shift-equivariance such as tokenization self-attention patch merging and positional encoding. With our proposed modules we achieve perfect circular shift-equivariance across four prominent ViT architectures: Swin SwinV2 CvT and MViTv2. Additionally we leverage our design to further enhance consistency under standard shifts. We evaluate our adaptive ViT models on image classification and semantic segmentation tasks. Our models achieve competitive performance across three diverse datasets showcasing perfect (100%) circular shift consistency while improving standard shift consistency.
****
SpikeNeRF: Learning Neural Radiance Fields from Continuous Spike Stream
Lin Zhu, Kangmin Jia, Yifan Zhao, Yunshan Qi, Lizhi Wang, Hua Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6285-6295
Spike cameras leveraging spike-based integration sampling and high temporal resolution offer distinct advantages over standard cameras. However existing approaches reliant on spike cameras often assume optimal illumination a condition frequently unmet in real-world scenarios. To address this we introduce SpikeNeRF the first work that derives a NeRF-based volumetric scene representation from spike camera data. Our approach leverages NeRF's multi-view consistency to establish robust self-supervision effectively eliminating erroneous measurements and uncovering coherent structures within exceedingly noisy input amidst diverse real-world illumination scenarios. The framework comprises two core elements: a spike generation model incorporating an integrate-and-fire neuron layer and parameters accounting for non-idealities such as threshold variation and a spike rendering loss capable of generalizing across varying illumination conditions. We describe how to effectively optimize neural radiance fields to render photorealistic novel views from the novel continuous spike stream demonstrating advantages over other vision sensors in certain scenes. Empirical evaluations conducted on both real and novel realistically simulated sequences affirm the efficacy of our methodology. The dataset and source code are released at https://github.com/BIT-Vision/SpikeNeRF.
****
Action Scene Graphs for Long-Form Understanding of Egocentric Videos
Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, Giovanni Maria Farinella; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18622-18632
We present Egocentric Action Scene Graphs (EASGs) a new representation for long-form understanding of egocentric videos. EASGs extend standard manually-annotated representations of egocentric videos such as verb-noun action labels by providing a temporally evolving graph-based description of the actions performed by the camera wearer including interacted objects their relationships and how actions unfold in time. Through a novel annotation procedure we extend the Ego4D dataset adding manually labeled Egocentric Action Scene Graphs which offer a rich set of annotations for long-from egocentric video understanding. We hence define the EASG generation task and provide a baseline approach establishing preliminary benchmarks. Experiments on two downstream tasks action anticipation and activity

summarization highlight the effectiveness of EASGs for long-form egocentric vide
o understanding. We will release the dataset and code to replicate experiments a
nd annotations.
****
A Semi-supervised Nighttime Dehazing Baseline with Spatial-Frequency Aware and R
ealistic Brightness Constraint
Xiaofeng Cong, Jie Gui, Jing Zhang, Junming Hou, Hao Shen; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2
631-2640
Existing research based on deep learning has extensively explored the problem of
 daytime image dehazing. However few studies have considered the characteristics
 of nighttime hazy scenes. There are two distinctions between nighttime and dayt
ime haze. First there may be multiple active colored light sources with lower il
lumination intensity in nighttime scenes which may cause haze glow and noise wit
h localized coupled and frequency inconsistent characteristics. Second due to th
e domain discrepancy between simulated and real-world data unrealistic brightnes
s may occur when applying a dehazing model trained on simulated data to real-wor
ld data. To address the above two issues we propose a semi-supervised model for
real-world nighttime dehazing. First the spatial attention and frequency spectru
m filtering are implemented as a spatial-frequency domain information interactio
n module to handle the first issue. Second a pseudo-label-based retraining strat
egy and a local window-based brightness loss for semi-supervised training proces
s is designed to suppress haze and glow while achieving realistic brightness. Ex
periments on public benchmarks validate the effectiveness of the proposed method
 and its superiority over state-of-the-art methods. The source code and Suppleme
ntary Materials are placed in the https://github.com/Xiaofeng-life/SFSNiD.
****
De-confounded Data-free Knowledge Distillation for Handling Distribution Shifts
Yuzheng Wang, Dingkang Yang, Zhaoyu Chen, Yang Liu, Siao Liu, Wenqiang Zhang, Li
hua Zhang, Lizhe Qi; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2024, pp. 12615-12625
Data-Free Knowledge Distillation (DFKD) is a promising task to train high-perfor
mance small models to enhance actual deployment without relying on the original
training data. Existing methods commonly avoid relying on private data by utiliz
ing synthetic or sampled data. However a long-overlooked issue is that the sever
e distribution shifts between their substitution and original data which manifes
ts as huge differences in the quality of images and class proportions. The harmf
ul shifts are essentially the confounder that significantly causes performance b
ottlenecks. To tackle the issue this paper proposes a novel perspective with cau
sal inference to disentangle the student models from the impact of such shifts.
By designing a customized causal graph we first reveal the causalities among the
 variables in the DFKD task. Subsequently we propose a Knowledge Distillation Ca
usal Intervention (KDCI) framework based on the backdoor adjustment to de-confou
nd the confounder. KDCI can be flexibly combined with most existing state-of-the
-art baselines. Experiments in combination with six representative DFKD methods
demonstrate the effectiveness of our KDCI which can obviously help existing meth
ods under almost all settings e.g. improving the baseline by up to 15.54% accura
cy on the CIFAR-100 dataset.
****
Fine-Grained Bipartite Concept Factorization for Clustering
Chong Peng, Pengfei Zhang, Yongyong Chen, Zhao Kang, Chenglizhao Chen, Qiang Che
ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2024, pp. 26264-26274
In this paper we propose a novel concept factorization method that seeks factor
matrices using a cross-order positive semi-definite neighbor graph which provide
s comprehensive and complementary neighbor information of the data. The factor m
atrices are learned with bipartite graph partitioning which exploits explicit cl
uster structure of the data and is more geared towards clustering application. W
e develop an effective and efficient optimization algorithm for our method and p
rovide elegant theoretical results about the convergence. Extensive experimental

results confirm the effectiveness of the proposed method.
****
Siamese Learning with Joint Alignment and Regression for Weakly-Supervised Video Paragraph Grounding
Chaolei Tan, Jianhuang Lai, Wei-Shi Zheng, Jian-Fang Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13569-13580
Video Paragraph Grounding (VPG) is an emerging task in video-language understanding which aims at localizing multiple sentences with semantic relations and temporal order from an untrimmed video. However existing VPG approaches are heavily reliant on a considerable number of temporal labels that are laborious and time-consuming to acquire. In this work we introduce and explore Weakly-Supervised Video Paragraph Grounding (WSVPG) to eliminate the need of temporal annotations. Different from previous weakly-supervised grounding frameworks based on multiple instance learning or reconstruction learning for two-stage candidate ranking we propose a novel siamese learning framework that jointly learns the cross-modal feature alignment and temporal coordinate regression without timestamp labels to achieve concise one-stage localization for WSVPG. Specifically we devise a Siamese Grounding TRansformer (SiamGTR) consisting of two weight-sharing branches for learning complementary supervision. An Augmentation Branch is utilized for directly regressing the temporal boundaries of a complete paragraph within a pseudo video and an Inference Branch is designed to capture the order-guided feature correspondence for localizing multiple sentences in a normal video. We demonstrate by extensive experiments that our paradigm has superior practicability and flexibility to achieve efficient weakly-supervised or semi-supervised learning outperforming state-of-the-art methods trained with the same or stronger supervision.
****
Language-Driven Anchors for Zero-Shot Adversarial Robustness
Xiao Li, Wei Zhang, Yining Liu, Zhanhao Hu, Bo Zhang, Xiaolin Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24686-24695
Deep Neural Networks (DNNs) are known to be susceptible to adversarial attacks. Previous researches mainly focus on improving adversarial robustness in the fully supervised setting leaving the challenging domain of zero-shot adversarial robustness an open question. In this work we investigate this domain by leveraging the recent advances in large vision-language models such as CLIP to introduce zero-shot adversarial robustness to DNNs. We propose LAAT a Language-driven Anchor-based Adversarial Training strategy. LAAT utilizes the features of a text encoder for each category as fixed anchors (normalized feature embeddings) for each category which are then employed for adversarial training. By leveraging the semantic consistency of the text encoders LAAT aims to enhance the adversarial robustness of the image model on novel categories. However naively using text encoders leads to poor results. Through analysis we identified the issue to be the high cosine similarity between text encoders. We then design an expansion algorithm and an alignment cross-entropy loss to alleviate the problem. Our experimental results demonstrated that LAAT significantly improves zero-shot adversarial robustness over state-of-the-art methods. LAAT has the potential to enhance adversarial robustness by large-scale multimodal models especially when labeled data is unavailable during training.
****
Deep Equilibrium Diffusion Restoration with Parallel Sampling
Jiezhang Cao, Yue Shi, Kai Zhang, Yulun Zhang, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2824-2834
Diffusion model-based image restoration (IR) aims to use diffusion models to recover high-quality (HQ) images from degraded images achieving promising performance. Due to the inherent property of diffusion models most existing methods need long serial sampling chains to restore HQ images step-by-step resulting in expensive sampling time and high computation costs. Moreover such long sampling chains hinder understanding the relationship between inputs and restoration results s

ince it is hard to compute the gradients in the whole chains. In this work we ai
m to rethink the diffusion model-based IR models through a different perspective
 i.e. a deep equilibrium (DEQ) fixed point system called DeqIR. Specifically we
derive an analytical solution by modeling the entire sampling chain in these IR
models as a joint multivariate fixed point system. Based on the analytical solut
ion we can conduct parallel sampling and restore HQ images without training. Fur
thermore we compute fast gradients via DEQ inversion and found that initializati
on optimization can boost image quality and control the generation direction. Ex
tensive experiments on benchmarks demonstrate the effectiveness of our method on
 typical IR tasks and real-world settings.
****
LEOD: Label-Efficient Object Detection for Event Cameras
Ziyi Wu, Mathias Gehrig, Qing Lyu, Xudong Liu, Igor Gilitschenski; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 16933-16943
Object detection with event cameras benefits from the sensor's low latency and h
igh dynamic range. However it is costly to fully label event streams for supervi
sed training due to their high temporal resolution. To reduce this cost we prese
nt LEOD the first method for label-efficient event-based detection. Our approach
 unifies weakly- and semi-supervised object detection with a self-training mecha
nism. We first utilize a detector pre-trained on limited labels to produce pseud
o ground truth on unlabeled events. Then the detector is re-trained with both re
al and generated labels. Leveraging the temporal consistency of events we run bi
-directional inference and apply tracking-based post-processing to enhance the q
uality of pseudo labels. To stabilize training against label noise we further de
sign a soft anchor assignment strategy. We introduce new experimental protocols
to evaluate the task of label-efficient event-based detection on Gen1 and 1Mpx d
atasets. LEOD consistently outperforms supervised baselines across various label
ing ratios. For example on Gen1 it improves mAP by 8.6% and 7.8% for RVT-S train
ed with 1% and 2% labels. On 1Mpx RVT-S with 10% labels even surpasses its fully
-supervised counterpart using 100% labels. LEOD maintains its effectiveness even
 when all labeled data are available reaching new state-of-the-art results. Fina
lly we show that our method readily scales to improve larger detectors as well.
Code is released at https://github.com/Wuziyi616/LEOD.
****
Morphological Prototyping for Unsupervised Slide Representation Learning in Comp
utational Pathology
Andrew H. Song, Richard J. Chen, Tong Ding, Drew F.K. Williamson, Guillaume Jaum
e, Faisal Mahmood; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2024, pp. 11566-11578
Representation learning of pathology whole-slide images (WSIs) has been has prim
arily relied on weak supervision with Multiple Instance Learning (MIL). However
the slide representations resulting from this approach are highly tailored to sp
ecific clinical tasks which limits their expressivity and generalization particu
larly in scenarios with limited data. Instead we hypothesize that morphological
redundancy in tissue can be leveraged to build a task-agnostic slide representat
ion in an unsupervised fashion. To this end we introduce PANTHER a prototype-bas
ed approach rooted in the Gaussian mixture model that summarizes the set of WSI
patches into a much smaller set of morphological prototypes. Specifically each p
atch is assumed to have been generated from a mixture distribution where each mi
xture component represents a morphological exemplar. Utilizing the estimated mix
ture parameters we then construct a compact slide representation that can be rea
dily used for a wide range of downstream tasks. By performing an extensive evalu
ation of PANTHER on subtyping and survival tasks using 13 datasets we show that
1) PANTHER outperforms or is on par with supervised MIL baselines and 2) the ana
lysis of morphological prototypes brings new qualitative and quantitative insigh
ts into model interpretability. The code is available at https://github.com/mahm
oodlab/Panther.
****
Fooling Polarization-Based Vision using Locally Controllable Polarizing Projecti

on
Zhuoxiao Li, Zhihang Zhong, Shohei Nobuhara, Ko Nishino, Yinqiang Zheng; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 24706-24715
Polarization is a fundamental property of light that encodes abundant informatio
n regarding surface shape material illumination and viewing geometry. The comput
er vision community has witnessed a blossom of polarization-based vision applica
tions such as reflection removal shape-from-polarization (SfP) transparent objec
t segmentation and color constancy partially due to the emergence of single-chip
 mono/color polarization sensors that make polarization data acquisition easier
than ever. However is polarization-based vision vulnerable to adversarial attack
s? If so is that possible to realize these adversarial attacks in the physical w
orld without being perceived by human eyes? In this paper we warn the community
of the vulnerability of polarization-based vision which can be more serious than
 RGB-based vision. By adapting a commercial LCD projector we achieve locally con
trollable polarizing projection which is successfully utilized to fool state-of-
the-art polarization-based vision algorithms for glass segmentation and SfP. Com
pared with existing physical attacks on RGB-based vision which always suffer fro
m the trade-off between attack efficacy and eye conceivability the adversarial a
ttackers based on polarizing projection are contact-free and visually impercepti
ble since naked human eyes can rarely perceive the difference of viciously manip
ulated polarizing light and ordinary illumination. This poses unprecedented risk
s on polarization-based vision for which due attentions should be paid and count
er measures be considered.
****
Dense Optical Tracking: Connecting the Dots
Guillaume Le Moing, Jean Ponce, Cordelia Schmid; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19187-19197
Recent approaches to point tracking are able to recover the trajectory of any sc
ene point through a large portion of a video despite the presence of occlusions.
 They are however too slow in practice to track every point observed in a single
 frame in a reasonable amount of time. This paper introduces DOT a novel simple
and efficient method for solving this problem. It first extracts a small set of
tracks from key regions at motion boundaries using an off-the-shelf point tracki
ng algorithm. Given source and target frames DOT then computes rough initial est
imates of a dense flow field and visibility mask through nearest-neighbor interp
olation before refining them using a learnable optical flow estimator that expli
citly handles occlusions and can be trained on synthetic data with ground-truth
correspondences. We show that DOT is significantly more accurate than current op
tical flow techniques outperforms sophisticated "universal" trackers like OmniMo
tion and is on par with or better than the best point tracking algorithms like C
oTracker while being at least two orders of magnitude faster. Quantitative and q
ualitative experiments with synthetic and real videos validate the promise of th
e proposed approach. Code data and videos showcasing the capabilities of our app
roach are available in the project webpage: https://16lemoing.github.io/dot .
****
A Stealthy Wrongdoer: Feature-Oriented Reconstruction Attack against Split Learn
ing
Xiaoyang Xu, Mengda Yang, Wenzhe Yi, Ziang Li, Juan Wang, Hongxin Hu, Yong Zhuan
g, Yaxin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2024, pp. 12130-12139
Split Learning (SL) is a distributed learning framework renowned for its privacy
-preserving features and minimal computational requirements. Previous research c
onsistently highlights the potential privacy breaches in SL systems by server ad
versaries reconstructing training data. However these studies often rely on stro
ng assumptions or compromise system utility to enhance attack performance. This
paper introduces a new semi-honest Data Reconstruction Attack on SL named Featur
e-Oriented Reconstruction Attack (FORA). In contrast to prior works FORA relies
on limited prior knowledge specifically that the server utilizes auxiliary sampl
es from the public without knowing any client's private information. This allows

FORA to conduct the attack stealthily and achieve robust performance. The key v
ulnerability exploited by FORA is the revelation of the model representation pre
ference in the smashed data output by victim client. FORA constructs a substitut
e client through feature-level transfer learning aiming to closely mimic the vic
tim client's representation preference. Leveraging this substitute client the se
rver trains the attack model to effectively reconstruct private data. Extensive
experiments showcase FORA's superior performance compared to state-of-the-art me
thods. Furthermore the paper systematically evaluates the proposed method's appl
icability across diverse settings and advanced defense strategies.
****
DiffAM: Diffusion-based Adversarial Makeup Transfer for Facial Privacy Protectio
n
Yuhao Sun, Lingyun Yu, Hongtao Xie, Jiaming Li, Yongdong Zhang; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 24584-24594
With the rapid development of face recognition (FR) systems the privacy of face
images on social media is facing severe challenges due to the abuse of unauthori
zed FR systems. Some studies utilize adversarial attack techniques to defend aga
inst malicious FR systems by generating adversarial examples. However the genera
ted adversarial examples i.e. the protected face images tend to suffer from subp
ar visual quality and low transferability. In this paper we propose a novel face
protection approach dubbed DiffAM which leverages the powerful generative abili
ty of diffusion models to generate high-quality protected face images with adver
sarial makeup transferred from reference images. To be specific we first introdu
ce a makeup removal module to generate non-makeup images utilizing a fine-tuned
diffusion model with guidance of textual prompts in CLIP space. As the inverse p
rocess of makeup transfer makeup removal can make it easier to establish the det
erministic relationship between makeup domain and non-makeup domain regardless o
f elaborate text prompts. Then with this relationship a CLIP-based makeup loss a
long with an ensemble attack strategy is introduced to jointly guide the directi
on of adversarial makeup domain achieving the generation of protected face image
s with natural-looking makeup and high black-box transferability. Extensive expe
riments demonstrate that DiffAM achieves higher visual quality and attack succes
s rates with a gain of 12.98% under black-box setting compared with the state of
the arts. The code will be available at https://github.com/HansSunY/DiffAM.
****
SlowFormer: Adversarial Attack on Compute and Energy Consumption of Efficient Vi
sion Transformers
K L Navaneet, Soroush Abbasi Koohpayegani, Essam Sleiman, Hamed Pirsiavash; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 24786-24797
Recently there has been a lot of progress in reducing the computation of deep mo
dels at inference time. These methods can reduce both the computational needs an
d power usage of deep models. Some of these approaches adaptively scale the comp
ute based on the input instance. We show that such models can be vulnerable to a
universal adversarial patch attack where the attacker optimizes for a patch tha
t when pasted on any image can increase the compute and power consumption of the
model. We run experiments with three different efficient vision transformer met
hods showing that in some cases the attacker can increase the computation to the
maximum possible level by simply pasting a patch that occupies only 8% of the i
mage area. We also show that a standard adversarial training defense method can
reduce some of the attack's success. We believe adaptive efficient methods will
be necessary for the future to lower the power usage of expensive deep models so
we hope our paper encourages the community to study the robustness of these met
hods and develop better defense methods for the proposed attack. Code is availab
le at: https://github.com/UCDvision/SlowFormer.
****
TULIP: Transformer for Upsampling of LiDAR Point Clouds
Bin Yang, Patrick Pfreundschuh, Roland Siegwart, Marco Hutter, Peyman Moghadam,
Vaishakh Patil; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa

ttern Recognition (CVPR), 2024, pp. 15354-15364
LiDAR Upsampling is a challenging task for the perception systems of robots and autonomous vehicles due to the sparse and irregular structure of large-scale scene contexts. Recent works propose to solve this problem by converting LiDAR data from 3D Euclidean space into an image super-resolution problem in 2D image space. Although their methods can generate high-resolution range images with fine-grained details the resulting 3D point clouds often blur out details and predict invalid points. In this paper we propose TULIP a new method to reconstruct high-resolution LiDAR point clouds from low-resolution LiDAR input. We also follow a range image-based approach but specifically modify the patch and window geometries of a Swin-Transformer-based network to better fit the characteristics of range images. We conducted several experiments on three public real-world and simulated datasets. TULIP outperforms state-of-the-art methods in all relevant metrics and generates robust and more realistic point clouds than prior works.
****
How to Configure Good In-Context Sequence for Visual Question Answering
Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, Xu Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26710-26720
Inspired by the success of Large Language Models in dealing with new tasks via In-Context Learning (ICL) in NLP researchers have also developed Large Vision-Language Models (LVLMs) with ICL capabilities. However when implementing ICL using these LVLMs researchers usually resort to the simplest way like random sampling to configure the in-context sequence thus leading to sub-optimal results. To enhance the ICL performance in this study we use Visual Question Answering (VQA) as case study to explore diverse in-context configurations to find the powerful ones. Additionally through observing the changes of the LVLM outputs by altering the in-context sequence we gain insights into the inner properties of LVLMs improving our understanding of them. Specifically to explore in-context configurations we design diverse retrieval methods and employ different strategies to manipulate the retrieved demonstrations. Through exhaustive experiments on three VQA datasets: VQAv2 VizWiz and OK-VQA we uncover three important inner properties of the applied LVLM and demonstrate which strategies can consistently improve the ICL VQA performance. Our code is provided in: https: //github.com/GaryJiajia/OFv2_ICL_VQA.
****
Gaussian Shell Maps for Efficient 3D Human Generation
Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9441-9451
Efficient generation of 3D digital humans is important in several industries including virtual reality social media and cinematic production. 3D generative adversarial networks (GANs) have demonstrated state-of-the-art (SOTA) quality and diversity for generated assets. Current 3D GAN architectures however typically rely on volume representations which are slow to render thereby hampering the GAN training and requiring multi-view-inconsistent 2D upsamplers. Here we introduce Gaussian Shell Maps (GSMs) as a framework that connects SOTA generator network architectures with emerging 3D Gaussian rendering primitives using an articulable multi shell--based scaffold. In this setting a CNN generates a 3D texture stack with features that are mapped to the shells. The latter represent inflated and deflated versions of a template surface of a digital human in a canonical body pose. Instead of rasterizing the shells directly we sample 3D Gaussians on the shells whose attributes are encoded in the texture features. These Gaussians are efficiently and differentiably rendered. The ability to articulate the shells is important during GAN training and at inference time to deform a body into arbitrary user-defined poses. Our efficient rendering scheme bypasses the need for view-inconsistent upsamplers and achieves high-quality multi-view consistent renderings at a native resolution of 512 x512 pixels. We demonstrate that GSMs successfully generate 3D humans when trained on single-view datasets including SHHQ and DeepFashion.

****
Defense Against Adversarial Attacks on No-Reference Image Quality Models with Gr adient Norm Regularization
Yujia Liu, Chenxi Yang, Dingquan Li, Jianhao Ding, Tingting Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25554-25563
The task of No-Reference Image Quality Assessment (NR-IQA) is to estimate the quality score of an input image without additional information. NR-IQA models play a crucial role in the media industry aiding in performance evaluation and optimization guidance. However these models are found to be vulnerable to adversarial attacks which introduce imperceptible perturbations to input images resulting in significant changes in predicted scores. In this paper we propose a defense method to mitigate the variability in predicted scores caused by small perturbations thus enhancing the adversarial robustness of NR-IQA models. To be specific we present theoretical evidence showing that the extent of score changes is related to the l_1 norm of the gradient of the predicted score with respect to the input image when adversarial perturbations are l_inf-bounded. Building on this theoretical foundation we propose a norm regularization training strategy aimed at reducing the l_1 norm of the gradient thereby boosting the adversarial robustness of NR-IQA models. Experiments conducted on four NR-IQA baseline models demonstrate the effectiveness of our strategy in reducing score changes in the presence of adversarial attacks. To the best of our knowledge this work marks the first attempt to defend against adversarial attacks on NR-IQA models. Our study offers valuable insights into the adversarial robustness of NR-IQA models and provides a foundation for future research in this area.
****
TACO: Benchmarking Generalizable Bimanual Tool-ACtion-Object Understanding
Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, Li Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21740-21751
Humans commonly work with multiple objects in daily life and can intuitively transfer manipulation skills to novel objects by understanding object functional regularities. However existing technical approaches for analyzing and synthesizing hand-object manipulation are mostly limited to handling a single hand and object due to the lack of data support. To address this we construct TACO an extensive bimanual hand-object-interaction dataset spanning a large variety of tool-action-object compositions for daily human activities. TACO contains 2.5K motion sequences paired with third-person and egocentric views precise hand-object 3D meshes and action labels. To rapidly expand the data scale we present a fully automatic data acquisition pipeline combining multi-view sensing with an optical motion capture system. With the vast research fields provided by TACO we benchmark three generalizable hand-object-interaction tasks: compositional action recognition generalizable hand-object motion forecasting and cooperative grasp synthesis. Extensive experiments reveal new insights challenges and opportunities for advancing the studies of generalizable hand-object motion analysis and synthesis. Our data and code are available at https://taco2024.github.io.
****
MoST: Motion Style Transformer Between Diverse Action Contents
Boeun Kim, Jungho Kim, Hyung Jin Chang, Jin Young Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1705-1714
While existing motion style transfer methods are effective between two motions with identical content their performance significantly diminishes when transferring style between motions with different contents. This challenge lies in the lack of clear separation between content and style of a motion. To tackle this challenge we propose a novel motion style transformer that effectively disentangles style from content and generates a plausible motion with transferred style from a source motion. Our distinctive approach to achieving the goal of disentanglement is twofold: (1) a new architecture for motion style transformer with 'part-attentive style modulator across body parts' and 'Siamese encoders that encode sty

le and content features separately'; (2) style disentanglement loss. Our method outperforms existing methods and demonstrates exceptionally high quality particularly in motion pairs with different contents without the need for heuristic post-processing. Codes are available at https://github.com/Boeun-Kim/MoST.
****
Prompting Hard or Hardly Prompting: Prompt Inversion for Text-to-Image Diffusion Models
Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, Leonid Sigal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6808-6817
The quality of the prompts provided to text-to-image diffusion models determines how faithful the generated content is to the user's intent often requiring `prompt engineering'. To harness visual concepts from target images without prompt engineering current approaches largely rely on embedding inversion by optimizing and then mapping them to pseudo-tokens. However working with such high-dimensional vector representations is challenging because they lack semantics and interpretability and only allow simple vector operations when using them. Instead this work focuses on inverting the diffusion model to obtain interpretable language prompts directly. The challenge of doing this lies in the fact that the resulting optimization problem is fundamentally discrete and the space of prompts is exponentially large; this makes using standard optimization techniques such as stochastic gradient descent difficult. To this end we utilize a delayed projection scheme to optimize for prompts representative of the vocabulary space in the model. Further we leverage the findings that different timesteps of the diffusion process cater to different levels of detail in an image. The later noisy timesteps of the forward diffusion process correspond to the semantic information and therefore prompt inversion in this range provides tokens representative of the image semantics. We show that our approach can identify semantically interpretable and meaningful prompts for a target image which can be used to synthesize diverse images with similar content. We further illustrate the application of the optimized prompts in evolutionary image generation and concept removal.
****
Unmixing Before Fusion: A Generalized Paradigm for Multi-Source-based Hyperspectral Image Synthesis
Yang Yu, Erting Pan, Xinya Wang, Yuheng Wu, Xiaoguang Mei, Jiayi Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9297-9306
In the realm of AI data serves as a pivotal resource. Real-world hyperspectral images (HSIs) bearing wide spectral characteristics are particularly valuable. However the acquisition of HSIs is always costly and time-intensive resulting in a severe data-thirsty issue in HSI research and applications. Current solutions have not been able to generate a sufficient volume of diverse and reliable synthetic HSIs. To this end our study formulates a novel generalized paradigm for HSI synthesis i.e. unmixing before fusion that initiates with unmixing across multi-source data and follows by fusion-based synthesis. By integrating unmixing this work maps unpaired HSI and RGB data to a low-dimensional abundance space greatly alleviating the difficulty of generating high-dimensional samples. Moreover incorporating abundances inferred from unpaired RGB images into generative models allows for cost-effective supplementation of various realistic spatial distributions in abundance synthesis. Our proposed paradigm can be instrumental with a series of deep generative models filling a significant gap in the field and enabling the generation of vast high-quality HSI samples for large-scale downstream tasks. Extension experiments on downstream tasks demonstrate the effectiveness of synthesized HSIs. The code is available at: HSI-Synthesis.github.io.
****
AlignMiF: Geometry-Aligned Multimodal Implicit Field for LiDAR-Camera Joint Synthesis
Tang Tao, Guangrun Wang, Yixing Lao, Peng Chen, Jie Liu, Liang Lin, Kaicheng Yu, Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21230-21240

Neural implicit fields have been a de facto standard in novel view synthesis. Re
cently there exist some methods exploring fusing multiple modalities within a si
ngle field aiming to share implicit features from different modalities to enhanc
e reconstruction performance. However these modalities often exhibit misaligned
behaviors: optimizing for one modality such as LiDAR can adversely affect anothe
r like camera performance and vice versa. In this work we conduct comprehensive
analyses on the multimodal implicit field of LiDAR-camera joint synthesis reveal
ing the underlying issue lies in the misalignment of different sensors. Furtherm
ore we introduce AlignMiF a geometrically aligned multimodal implicit field with
 two proposed modules: Geometry-Aware Alignment (GAA) and Shared Geometry Initia
lization (SGI). These modules effectively align the coarse geometry across diffe
rent modalities significantly enhancing the fusion process between LiDAR and cam
era data. Through extensive experiments across various datasets and scenes we de
monstrate the effectiveness of our approach in facilitating better interaction b
etween LiDAR and camera modalities within a unified neural field. Specifically o
ur proposed AlignMiF achieves remarkable improvement over recent implicit fusion
 methods (+2.01 and +3.11 image PSNR on the KITTI-360 and Waymo datasets) and co
nsistently surpasses single modality performance (13.8% and 14.2% reduction in L
iDAR Chamfer Distance on the respective datasets).
****
CoDi: Conditional Diffusion Distillation for Higher-Fidelity and Faster Image Ge
neration
Kangfu Mei, Mauricio Delbracio, Hossein Talebi, Zhengzhong Tu, Vishal M. Patel,
Peyman Milanfar; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2024, pp. 9048-9058
Large generative diffusion models have revolutionized text-to-image generation a
nd offer immense potential for conditional generation tasks such as image enhanc
ement restoration editing and compositing. However their widespread adoption is
hindered by the high computational cost which limits their real-time application
. To address this challenge we introduce a novel method dubbed CoDi that adapts
a pre-trained latent diffusion model to accept additional image conditioning inp
uts while significantly reducing the sampling steps required to achieve high-qua
lity results. Our method can leverage architectures such as ControlNet to incorp
orate conditioning inputs without compromising the model's prior knowledge gaine
d during large scale pre-training. Additionally a conditional consistency loss e
nforces consistent predictions across diffusion steps effectively compelling the
 model to generate high-quality images with conditions in a few steps. Our condi
tional-task learning and distillation approach outperforms previous distillation
 methods achieving a new state-of-the-art in producing high-quality images with
very few steps (e.g. 1-4) across multiple tasks including super-resolution text-
guided image editing and depth-to-image generation.
****
Improving Unsupervised Hierarchical Representation with Reinforcement Learning
Ruyi An, Yewen Li, Xu He, Pengjie Gu, Mengchen Zhao, Dong Li, Jianye Hao, Chaoji
e Wang, Bo An, Mingyuan Zhou; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2024, pp. 22946-22956
Learning representations to capture the very fundamental understanding of the wo
rld is a key challenge in machine learning. The hierarchical structure of explan
atory factors hidden in data is such a general representation and could be poten
tially achieved with a hierarchical VAE. However training a hierarchical VAE alw
ays suffers from the "posterior collapse" where the data information is hard to
propagate to the higher-level latent variables hence resulting in a bad hierarch
ical representation. To address this issue we first analyze the shortcomings of
existing methods for mitigating the "posterior collapse" from an information the
ory perspective then highlight the necessity of regularization for explicitly pr
opagating data information to higher-level latent variables while maintaining th
e dependency between different levels. This naturally leads to formulating the i
nference of the hierarchical latent representation as a sequential decision proc
ess which could benefit from applying reinforcement learning (RL). Aligning RL's
 objective with the regularization we first introduce a "skip-generative path" t

o acquire a reward for evaluating the information content of an inferred latent representation and then the developed Q-value function based on it could have a consistent optimization direction of the regularization. Finally policy gradient one of the typical RL methods is employed to train a hierarchical VAE without introducing a gradient estimator. Experimental results firmly support our analysis and demonstrate that our proposed method effectively mitigates the "posterior collapse" issue learns an informative hierarchy acquires explainable latent representations and significantly outperforms other hierarchical VAE-based methods in downstream tasks.
****
HPL-ESS: Hybrid Pseudo-Labeling for Unsupervised Event-based Semantic Segmentation
Linglin Jing, Yiming Ding, Yunpeng Gao, Zhigang Wang, Xu Yan, Dong Wang, Gerald Schaefer, Hui Fang, Bin Zhao, Xuelong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23128-23137
Event-based semantic segmentation has gained popularity due to its capability to deal with scenarios under high-speed motion and extreme lighting conditions which cannot be addressed by conventional RGB cameras. Since it is hard to annotate event data previous approaches rely on event-to-image reconstruction to obtain pseudo labels for training. However this will inevitably introduce noise and learning from noisy pseudo labels especially when generated from a single source may reinforce the errors. This drawback is also called confirmation bias in pseudo-labeling. In this paper we propose a novel hybrid pseudo-labeling framework for unsupervised event-based semantic segmentation HPL-ESS to alleviate the influence of noisy pseudo labels. In particular we first employ a plain unsupervised domain adaptation framework as our baseline which can generate a set of pseudo labels through self-training. Then we incorporate offline event-to-image reconstruction into the framework and obtain another set of pseudo labels by predicting segmentation maps on the reconstructed images. A noisy label learning strategy is designed to mix the two sets of pseudo labels and enhance the quality. Moreover we propose a soft prototypical alignment module to further improve the consistency of target domain features. Extensive experiments show that our proposed method outperforms existing state-of-the-art methods by a large margin on the DSEC-Semantic dataset (+5.88% accuracy +10.32% mIoU) which even surpasses several supervised methods.
****
X-Adapter: Adding Universal Compatibility of Plugins for Upgraded Diffusion Model
Lingmin Ran, Xiaodong Cun, Jia-Wei Liu, Rui Zhao, Song Zijie, Xintao Wang, Jussi Keppo, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8775-8784
We introduce X-Adapter a universal upgrader to enable the pretrained plug-and-play modules (e.g. ControlNet LoRA) to work directly with the upgraded text-to-image diffusion model (e.g. SDXL) without further retraining. We achieve this goal by training an additional network to control the frozen upgraded model with the new text-image data pairs. In detail X-Adapter keeps a frozen copy of the old model to preserve the connectors of different plugins. Additionally X-Adapter adds trainable mapping layers that bridge the decoders from models of different versions for feature remapping. The remapped features will be used as guidance for the upgraded model. To enhance the guidance ability of X-Adapter we employ a -text training strategy for the upgraded model. After training we also introduce a two-stage denoising strategy to align the initial latents of X-Adapter and the upgraded model. Thanks to our strategies X-Adapter demonstrates universal compatibility with various plugins and also enables plugins of different versions to work together thereby expanding the functionalities of diffusion community. To verify the effectiveness of the proposed method we conduct extensive experiments and the results show that X-Adapter may facilitate wider application in the upgraded foundational diffusion model. Project page at: https://showlab.github.io/X-Adapter.
****

Towards General Robustness Verification of MaxPool-based Convolutional Neural Ne tworks via Tightening Linear Approximation
Yuan Xiao, Shiqing Ma, Juan Zhai, Chunrong Fang, Jinyuan Jia, Zhenyu Chen; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2024, pp. 24766-24775
The robustness of convolutional neural networks (CNNs) is vital to modern AI-dri ven systems. It can be quantified by formal verification by providing a certifie d lower bound within which any perturbation does not alter the original input's classification result. It is challenging due to nonlinear components such as Max Pool. At present many verification methods are sound but risk losing some precis ion to enhance efficiency and scalability and thus a certified lower bound is a crucial criterion for evaluating the performance of verification tools. In this paper we present MaxLin a robustness verifier for MaxPool-based CNNs with tight Linear approximation. By tightening the linear approximation of the MaxPool func tion we can certify larger certified lower bounds of CNNs. We evaluate MaxLin wi th open-sourced benchmarks including LeNet and networks trained on the MNIST CIF AR-10 and Tiny ImageNet datasets. The results show that MaxLin outperforms state -of-the-art tools with up to 110.60% improvement regarding the certified lower b ound and 5.13 X speedup for the same neural networks. Our code is available at h ttps://github.com/xiaoyuanpigo/maxlin.
****
BT-Adapter: Video Conversation is Feasible Without Video Instruction Tuning
Ruyang Liu, Chen Li, Yixiao Ge, Thomas H. Li, Ying Shan, Ge Li; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13658-13667
The recent progress in Large Language Models (LLM) has spurred various advanceme nts in image-language conversation agents while how to build a proficient video- based dialogue system is still under exploration. Considering the extensive scal e of LLM and visual backbone minimal GPU memory is left for facilitating effecti ve temporal modeling which is crucial for comprehending and providing feedback o n videos. To this end we propose Branching Temporal Adapter (BT-Adapter) a novel method for extending image-language pretrained models into the video domain. Sp ecifically BT-Adapter serves as a plug-and-use temporal modeling branch alongsid e the pretrained visual encoder which is tuned while keeping the backbone frozen . Just pretrained once BT-Adapter can be seamlessly integrated into all image co nversation models using this version of CLIP enabling video conversations withou t the need for video instructions. Besides we develop a unique asymmetric token masking strategy inside the branch with tailor-made training tasks for BT-Adapte r facilitating faster convergence and better results. Thanks to BT-Adapter we ar e able to empower existing multimodal dialogue models with strong video understa nding capabilities without incurring excessive GPU costs. Without bells and whis tles BT-Adapter achieves (1) state-of-the-art zero-shot results on various video tasks using thousands of fewer GPU hours. (2) better performance than current v ideo chatbots without any video instruction tuning. (3) state-of-the-art results of video chatting using video instruction tuning outperforming previous SOTAs b y a large margin. The code has been available at https://github.com/farewellthre e/BT-Adapter.
****
CADTalk: An Algorithm and Benchmark for Semantic Commenting of CAD Programs
Haocheng Yuan, Jing Xu, Hao Pan, Adrien Bousseau, Niloy J. Mitra, Changjian Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2024, pp. 3753-3762
CAD programs are a popular way to compactly encode shapes as a sequence of opera tions that are easy to parametrically modify. However without sufficient semanti c comments and structure such programs can be challenging to understand let alon e modify. We introduce the problem of semantic commenting CAD programs wherein t he goal is to segment the input program into code blocks corresponding to semant ically meaningful shape parts and assign a semantic label to each block. We solv e the problem by combining program parsing with visual-semantic analysis afforde d by recent advances in foundational language and vision models. Specifically by

executing the input programs we create shapes which we use to generate conditional photorealistic images to make use of semantic annotators for such images. We then distill the information across the images and link back to the original programs to semantically comment on them. Additionally we collected and annotated a benchmark dataset CADTalk consisting of 5288 machine-made programs and 45 human-made programs with ground truth semantic comments. We extensively evaluated our approach compared it to a GPT-based baseline and an open-set shape segmentation baseline and reported an 83.24% accuracy on the new CADTalk dataset. Code and data: https://enigma-li.github.io/CADTalk/.
****
Learning to Rematch Mismatched Pairs for Robust Cross-Modal Retrieval
Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26679-26688
Collecting well-matched multimedia datasets is crucial for training cross-modal retrieval models. However in real-world scenarios massive multimodal data are harvested from the Internet which inevitably contains Partially Mismatched Pairs (PMPs). Undoubtedly such semantical irrelevant data will remarkably harm the cross-modal retrieval performance. Previous efforts tend to mitigate this problem by estimating a soft correspondence to down-weight the contribution of PMPs. In this paper we aim to address this challenge from a new perspective: the potential semantic similarity among unpaired samples makes it possible to excavate useful knowledge from mismatched pairs. To achieve this we propose L2RM a general framework based on Optimal Transport (OT) that learns to rematch mismatched pairs. In detail L2RM aims to generate refined alignments by seeking a minimal-cost transport plan across different modalities. To formalize the rematching idea in OT first we propose a self-supervised cost function that automatically learns from explicit similarity-cost mapping relation. Second we present to model a partial OT problem while restricting the transport among false positives to further boost refined alignments. Extensive experiments on three benchmarks demonstrate our L2RM significantly improves the robustness against PMPs for existing models. The code is available at https://github.com/hhc1997/L2RM.
****
Generate Subgoal Images before Act: Unlocking the Chain-of-Thought Reasoning in Diffusion Model for Robot Manipulation with Multimodal Prompts
Fei Ni, Jianye Hao, Shiguang Wu, Longxin Kou, Jiashun Liu, Yan Zheng, Bin Wang, Yuzheng Zhuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13991-14000
Robotics agents often struggle to understand and follow the multi-modal prompts in complex manipulation scenes which are challenging to be sufficiently and accurately described by text alone. Moreover for long-horizon manipulation tasks the deviation from general instruction tends to accumulate if lack of intermediate guidance from high-level subgoals. For this we consider can we generate subgoal images before act to enhance the instruction following in long-horizon manipulation with multi-modal prompts? Inspired by the great success of diffusion model in image generation tasks we propose a novel hierarchical framework named as CoTDiffusion that incorporates diffusion model as a high-level planner to convert the general and multi-modal prompts into coherent visual subgoal plans which further guide the low-level policy model before action execution. We design a semantic alignment module that can anchor the progress of generated keyframes along a coherent generation chain unlocking the chain-of-thought reasoning ability of diffusion model. Additionally we propose bi-directional generation and frame concat mechanism to further enhance the fidelity of generated subgoal images and the accuracy of instruction following. The experiments cover various robotics manipulation scenarios including visual reasoning visual rearrange and visual constraints. CoTDiffusion achieves outstanding performance gain compared to the baselines without explicit subgoal generation which proves that a subgoal image is worth a thousand words of instruction.
****
Asymmetric Masked Distillation for Pre-Training Small Foundation Models

Zhiyu Zhao, Bingkun Huang, Sen Xing, Gangshan Wu, Yu Qiao, Limin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18516-18526
Self-supervised foundation models have shown great potential in computer vision thanks to the pre-training paradigm of masked autoencoding. Scale is a primary factor influencing the performance of these foundation models. However these large foundation models often result in high computational cost. This paper focuses on pre-training relatively small vision transformer models that could be efficiently adapted to downstream tasks. Specifically taking inspiration from knowledge distillation in model compression we propose a new asymmetric masked distillation (AMD) framework for pre-training relatively small models with autoencoding. The core of AMD is to devise an asymmetric masking strategy where the teacher model is enabled to see more context information with a lower masking ratio while the student model is still equipped with a high masking ratio. We design customized multi-layer feature alignment between the teacher encoder and student encoder to regularize the pre-training of student MAE. To demonstrate the effectiveness and versatility of AMD we apply it to both ImageMAE and VideoMAE for pre-training relatively small ViT models. AMD achieved 84.6% classification accuracy on IN1K using the ViT-B model. And AMD achieves 73.3% classification accuracy using the ViT-B model on the Something-in-Something V2 dataset a 3.7% improvement over the original ViT-B model from VideoMAE. We also transfer AMD pre-trained models to downstream tasks and obtain consistent performance improvement over the original masked autoencoding. The code and models are available at https://github.com/MCG-NJU/AMD.
****
Inversion-Free Image Editing with Language-Guided Diffusion Models
Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, Joyce Chai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9452-9461
Despite recent advances in inversion-based editing text-guided image manipulation remains challenging for diffusion models. The primary bottlenecks include 1) the time-consuming nature of the inversion process; 2) the struggle to balance consistency with accuracy; 3) the lack of compatibility with efficient consistency sampling methods used in consistency models. To address the above issues we start by asking ourselves if the inversion process can be eliminated for editing. We show that when the initial sample is known a special variance schedule reduces the denoising step to the same form as the multi-step consistency sampling. We name this Denoising Diffusion Consistent Model (DDCM) and note that it implies a virtual inversion strategy without explicit inversion in sampling. We further unify the attention control mechanisms in a tuning-free framework for text-guided editing. Combining them we present inversion-free editing (InfEdit) which allows for consistent and faithful editing for both rigid and non-rigid semantic changes catering to intricate modifications without compromising on the image's integrity and explicit inversion. Through extensive experiments InfEdit shows strong performance in various editing tasks and also maintains a seamless workflow (less than 3 seconds on one single A40) demonstrating the potential for real-time applications.
****
HumMUSS: Human Motion Understanding using State Space Models
Arnab Mondal, Stefano Alletto, Denis Tome; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2318-2330
Understanding human motion from video is essential for a range of applications including pose estimation mesh recovery and action recognition. While state-of-the-art methods predominantly rely on transformer-based architectures these approaches have limitations in practical scenarios. Transformers are slower when sequentially predicting on a continuous stream of frames in real-time and do not generalize to new frame rates. In light of these constraints we propose a novel attention-free spatiotemporal model for human motion understanding building upon recent advancements in state space models. Our model not only matches the performance of transformer-based models in various motion understanding tasks but also br

ings added benefits like adaptability to different video frame rates and enhanced training speed when working with longer sequence of keypoints. Moreover the proposed model supports both offline and real-time applications. For real-time sequential prediction our model is both memory efficient and several times faster than transformer-based approaches while maintaining their high accuracy.
****
MP5: A Multi-modal Open-ended Embodied System in Minecraft via Active Perception
Yiran Qin, Enshen Zhou, Qichang Liu, Zhenfei Yin, Lu Sheng, Ruimao Zhang, Yu Qiao, Jing Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16307-16316
It is a long-lasting goal to design an embodied system that can solve long-horizon open-world tasks in human-like ways. However existing approaches usually struggle with compound difficulties caused by the logic-aware decomposition and context-aware execution of these tasks. To this end we introduce MP5 an open-ended multimodal embodied system built upon the challenging Minecraft simulator which can decompose feasible sub-objectives design sophisticated situation-aware plans and perform embodied action control with frequent communication with a goal-conditioned active perception scheme. Specifically MP5 is developed on top of recent advances in Multimodal Large Language Models (MLLMs) and the system is modulated into functional modules that can be scheduled and collaborated to ultimately solve pre-defined context- and process-dependent tasks. Extensive experiments prove that MP5 can achieve a 22% success rate on difficult process-dependent tasks and a 91% success rate on tasks that heavily depend on the context. Moreover MP5 exhibits a remarkable ability to address many open-ended tasks that are entirely novel.
****
Uncovering What Why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly
Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, Xiaofeng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18793-18803
Video anomaly understanding (VAU) aims to automatically comprehend unusual occurrences in videos thereby enabling various applications such as traffic surveillance and industrial manufacturing. While existing VAU benchmarks primarily concentrate on anomaly detection and localization our focus is on more practicality prompting us to raise the following crucial questions: "what anomaly occurred?" "why did it happen?" and "how severe is this abnormal event?". In pursuit of these answers we present a comprehensive benchmark for Causation Understanding of Video Anomaly (CUVA). Specifically each instance of the proposed benchmark involves three sets of human annotations to indicate the "what" "why" and "how" of an anomaly including 1) anomaly type start and end times and event descriptions 2) natural language explanations for the cause of an anomaly and 3) free text reflecting the effect of the abnormality. In addition we also introduce MMEval a novel evaluation metric designed to better align with human preferences for CUVA facilitating the measurement of existing LLMs in comprehending the underlying cause and corresponding effect of video anomalies. Finally we propose a novel prompt-based method that can serve as a baseline approach for the challenging CUVA. We conduct extensive experiments to show the superiority of our evaluation metric and the prompt-based approach.
****
MiKASA: Multi-Key-Anchor & Scene-Aware Transformer for 3D Visual Grounding
Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, Didier Stricker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14131-14140
3D visual grounding involves matching natural language descriptions with their corresponding objects in 3D spaces. Existing methods often face challenges with accuracy in object recognition and struggle in interpreting complex linguistic queries particularly with descriptions that involve multiple anchors or are view-d

ependent. In response we present the MiKASA (Multi-Key-Anchor Scene-Aware) Transformer. Our novel end-to-end trained model integrates a self-attention-based scene-aware object encoder and an original multi-key-anchor technique enhancing object recognition accuracy and the understanding of spatial relationships. Furthermore MiKASA improves the explainability of decision-making facilitating error diagnosis. Our model achieves the highest overall accuracy in the Referit3D challenge for both the Sr3D and Nr3D datasets particularly excelling by a large margin in categories that require viewpoint-dependent descriptions.
****
ZePT: Zero-Shot Pan-Tumor Segmentation via Query-Disentangling and Self-Prompting
Yankai Jiang, Zhongzhen Huang, Rongzhao Zhang, Xiaofan Zhang, Shaoting Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11386-11397
The long-tailed distribution problem in medical image analysis reflects a high prevalence of common conditions and a low prevalence of rare ones which poses a significant challenge in developing a unified model capable of identifying rare or novel tumor categories not encountered during training. In this paper we propose a new Zero-shot Pan-Tumor segmentation framework (ZePT) based on query-disentangling and self-prompting to segment unseen tumor categories beyond the training set. ZePT disentangles the object queries into two subsets and trains them in two stages. Initially it learns a set of fundamental queries for organ segmentation through an object-aware feature grouping strategy which gathers organ-level visual features. Subsequently it refines the other set of advanced queries that focus on the auto-generated visual prompts for unseen tumor segmentation. Moreover we introduce query-knowledge alignment at the feature level to enhance each query's discriminative representation and generalizability. Extensive experiments on various tumor segmentation tasks demonstrate the performance superiority of ZePT which surpasses the previous counterparts and evidences the promising ability for zero-shot tumor segmentation in real-world settings.
****
Task-Driven Exploration: Decoupling and Inter-Task Feedback for Joint Moment Retrieval and Highlight Detection
Jin Yang, Ping Wei, Huan Li, Ziyang Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18308-18318
Video moment retrieval and highlight detection are two highly valuable tasks in video understanding but until recently they have been jointly studied. Although existing studies have made impressive advancement recently they predominantly follow the data-driven bottom-up paradigm. Such paradigm overlooks task-specific and inter-task effects resulting in poor model performance. In this paper we propose a novel task-driven top-down framework TaskWeave for joint moment retrieval and highlight detection. The framework introduces a task-decoupled unit to capture task-specific and common representations. To investigate the interplay between the two tasks we propose an inter-task feedback mechanism which transforms the results of one task as guiding masks to assist the other task. Different from existing methods we present a task-dependent joint loss function to optimize the model. Comprehensive experiments and in-depth ablation studies on QVHighlights TVSum and Charades-STA datasets corroborate the effectiveness and flexibility of the proposed framework. Codes are available at https://github.com/EdenGabriel/TaskWeave.
****
MobileCLIP: Fast Image-Text Models through Multi-Modal Reinforced Training
Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, Oncel Tuzel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15963-15974
Contrastive pre-training of image-text foundation models such as CLIP demonstrated excellent zero-shot performance and improved robustness on a wide range of downstream tasks. However these models utilize large transformer-based encoders with significant memory and latency overhead which pose challenges for deployment on mobile devices. In this work we introduce MobileCLIP - a new family of effici

ent image-text models optimized for runtime performance along with a novel and efficient training approach namely multi-modal reinforced training. The proposed training approach leverages knowledge transfer from an image captioning model and an ensemble of strong CLIP encoders to improve the accuracy of efficient models. Our approach avoids train-time compute overhead by storing the additional knowledge in a reinforced dataset. MobileCLIP sets a new state-of-the-art latency-accuracy tradeoff for zero-shot classification and retrieval tasks on several datasets. Our MobileCLIP-S2 variant is 2.3x faster while more accurate compared to previous best CLIP model based on ViT-B/16. We further demonstrate the effectiveness of our multi-modal reinforced training by training a CLIP model based on ViT-B/16 image backbone and achieving +2.9% average performance improvement on 38 evaluation benchmarks compared to the previous best. Moreover we show that the proposed approach achieves 10x-1000x improved learning efficiency when compared with non- reinforced CLIP training. Code and models are available at https://github.com/apple/ml-mobileclip
****
Drag Your Noise: Interactive Point-based Editing via Diffusion Semantic Propagation
Haofeng Liu, Chenshu Xu, Yifei Yang, Lihua Zeng, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6743-6752
Point-based interactive editing serves as an essential tool to complement the controllability of existing generative models. A concurrent work DragDiffusion updates the diffusion latent map in response to user inputs causing global latent map alterations. This results in imprecise preservation of the original content and unsuccessful editing due to gradient vanishing. In contrast we present DragNoise offering robust and accelerated editing without retracing the latent map. The core rationale of DragNoise lies in utilizing the predicted noise output of each U-Net as a semantic editor. This approach is grounded in two critical observations: firstly the bottleneck features of U-Net inherently possess semantically rich features ideal for interactive editing; secondly high-level semantics established early in the denoising process show minimal variation in subsequent stages. Leveraging these insights DragNoise edits diffusion semantics in a single denoising step and efficiently propagates these changes ensuring stability and efficiency in diffusion editing. Comparative experiments reveal that DragNoise achieves superior control and semantic retention reducing the optimization time by over 50% compared to DragDiffusion. Our codes are available at https://github.com/haofengl/DragNoise.
****
CDMAD: Class-Distribution-Mismatch-Aware Debiasing for Class-Imbalanced Semi-Supervised Learning
Hyuck Lee, Heeyoung Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23891-23900
Pseudo-label-based semi-supervised learning (SSL) algorithms trained on a class-imbalanced set face two cascading challenges: 1) Classifiers tend to be biased towards majority classes and 2) Biased pseudo-labels are used for training. It is difficult to appropriately re-balance the classifiers in SSL because the class distribution of an unlabeled set is often unknown and could be mismatched with that of a labeled set. We propose a novel class-imbalanced SSL algorithm called class-distribution-mismatch-aware debiasing (CDMAD). For each iteration of training CDMAD first assesses the classifier's biased degree towards each class by calculating the logits on an image without any patterns (e.g. solid color image) which can be considered irrelevant to the training set. CDMAD then refines biased pseudo-labels of the base SSL algorithm by ensuring the classifier's neutrality. CDMAD uses these refined pseudo-labels during the training of the base SSL algorithm to improve the quality of the representations. In the test phase CDMAD similarly refines biased class predictions on test samples. CDMAD can be seen as an extension of post-hoc logit adjustment to address a challenge of incorporating the unknown class distribution of the unlabeled set for re-balancing the biased classifier under class distribution mismatch. CDMAD ensures Fisher consistency f

or the balanced error. Extensive experiments verify the effectiveness of CDMAD.
****
VideoCon: Robust Video-Language Alignment via Contrast Captions
Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, Aditya Grover; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13927-13937
Despite being (pre)trained on a massive amount of data state-of-the-art video-language alignment models are not robust to semantically-plausible contrastive changes in the video captions. Our work addresses this by identifying a broad spectrum of contrast misalignments such as replacing entities actions and flipping event order which alignment models should be robust against. To this end we introduce the VideoCon a video-language alignment dataset constructed by a large language model that generates plausible contrast video captions and explanations for differences between original and contrast video captions. Then a generative video-language model is finetuned with VideoCon to assess video-language entailment and generate explanations. Our VideoCon-based alignment model significantly outperforms current models. It exhibits a 12-point increase in AUC for the video-language alignment task on human-generated contrast captions. Finally our model sets new state of the art zero-shot performance in temporally-extensive video-language tasks such as text-to-video retrieval (SSv2-Temporal) and video question answering (ATP-Hard). Moreover our model shows superior performance on novel videos and human-crafted captions and explanations.
****
PanoPose: Self-supervised Relative Pose Estimation for Panoramic Images
Diantao Tu, Hainan Cui, Xianwei Zheng, Shuhan Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20009-20018
Scaled relative pose estimation i.e. estimating relative rotation and scaled relative translation between two images has always been a major challenge in global Structure-from-Motion (SfM). This difficulty arises because the two-view relative translation computed by traditional geometric vision methods e.g. the five-point algorithm is scaleless. Many researchers have proposed diverse translation averaging methods to solve this problem. Instead of solving the problem in the motion averaging phase we focus on estimating scaled relative pose with the help of panoramic cameras and deep neural networks. In this paper a novel network namely PanoPose is proposed to estimate the relative motion in a fully self-supervised manner and a global SfM pipeline is built for panorama images. The proposed PanoPose comprises a depth-net and a pose-net with self-supervision achieved by reconstructing the reference image from its neighboring images based on the estimated depth and relative pose. To maintain precise pose estimation under large viewing angle differences we randomly rotate the panoramic images and pre-train the pose-net with images before and after the rotation. To enhance scale accuracy a fusion block is introduced to incorporate depth information into pose estimation. Extensive experiments on panoramic SfM datasets demonstrate the effectiveness of PanoPose compared with state-of-the-arts.
****
ContextSeg: Sketch Semantic Segmentation by Querying the Context with Attention
Jiawei Wang, Changjian Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3679-3688
Sketch semantic segmentation is a well-explored and pivotal problem in computer vision involving the assignment of predefined part labels to individual strokes. This paper presents ContextSeg - a simple yet highly effective approach to tackling this problem with two stages. In the first stage to better encode the shape and positional information of strokes we propose to predict an extra dense distance field in an autoencoder network to reinforce structural information learning. In the second stage we treat an entire stroke as a single entity and label a group of strokes within the same semantic part using an autoregressive Transformer with the default attention mechanism. By group-based labeling our method can fully leverage the context information when making decisions for the remaining groups of strokes. Our method achieves the best segmentation accuracy compared wi

th state-of-the-art approaches on two representative datasets and has been extensively evaluated demonstrating its superior performance. Additionally we offer insights into solving part imbalance in training data and the preliminary experiment on cross-category training which can inspire future research in this field.
****
Describing Differences in Image Sets with Natural Language
Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E. Gonzalez, Serena Yeung-Levy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24199-24208
How do two sets of images differ? Discerning set-level differences is crucial for understanding model behaviors and analyzing datasets yet manually sifting through thousands of images is impractical. To aid in this discovery process we explore the task of automatically describing the differences between two sets of images which we term Set Difference Captioning. This task takes in image sets \mathcal D _A and \mathcal D _B and outputs a description that is more often true on \mathcal D _A than \mathcal D _B. We outline a two-stage approach that first proposes candidate difference descriptions from image sets and then re-ranks the candidates by checking how well they can differentiate the two sets. We introduce VisDiff which first captions the images and prompts a language model to propose candidate descriptions then re-ranks these descriptions using CLIP. To evaluate VisDiff we collect VisDiffBench a dataset with 187 paired image sets with ground truth difference descriptions. We apply VisDiff to various domains such as comparing datasets (e.g. ImageNet vs. ImageNetV2) comparing classification models (e.g. zero-shot CLIP vs. supervised ResNet) characterizing differences between generative models (e.g. StableDiffusionV1 and V2) and discovering what makes images memorable. Using VisDiff we are able to find interesting and previously unknown differences in datasets and models demonstrating its utility in revealing nuanced insights.
****
Discovering and Mitigating Visual Biases through Keyword Explanation
Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, Jinwoo Shin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11082-11092
Addressing biases in computer vision models is crucial for real-world AI deployments. However mitigating visual biases is challenging due to their unexplainable nature often identified indirectly through visualization or sample statistics which necessitates additional human supervision for interpretation. To tackle this issue we propose the Bias-to-Text (B2T) framework which interprets visual biases as keywords. Specifically we extract common keywords from the captions of mispredicted images to identify potential biases in the model. We then validate these keywords by measuring their similarity to the mispredicted images using a vision-language scoring model. The keyword explanation form of visual bias offers several advantages such as a clear group naming for bias discovery and a natural extension for debiasing using these group names. Our experiments demonstrate that B2T can identify known biases such as gender bias in CelebA background bias in Waterbirds and distribution shifts in ImageNet-R/C. Additionally B2T uncovers novel biases in larger datasets such as Dollar Street and ImageNet. For example we discovered a contextual bias between \keyword bee and \keyword flower in ImageNet. We also highlight various applications of B2T keywords including debiased training CLIP prompting and model comparison.
****
Robust Emotion Recognition in Context Debiasing
Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, Lihua Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12447-12457
Context-aware emotion recognition (CAER) has recently boosted the practical applications of affective computing techniques in unconstrained environments. Mainstream CAER methods invariably extract ensemble representations from diverse contexts and subject-centred characteristics to perceive the target person's emotional state. Despite advancements the biggest challenge remains due to context bias

interference. The harmful bias forces the models to rely on spurious correlation s between background contexts and emotion labels in likelihood estimation causin g severe performance bottlenecks and confounding valuable context priors. In thi s paper we propose a counterfactual emotion inference (CLEF) framework to addres s the above issue. Specifically we first formulate a generalized causal graph to decouple the causal relationships among the variables in CAER. Following the ca usal graph CLEF introduces a non-invasive context branch to capture the adverse direct effect caused by the context bias. During the inference we eliminate the direct context effect from the total causal effect by comparing factual and coun terfactual outcomes resulting in bias mitigation and robust prediction. As a mod el-agnostic framework CLEF can be readily integrated into existing methods bring ing consistent performance gains.
****
Fully Geometric Panoramic Localization
Junho Kim, Jiwon Jeong, Young Min Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20827-20837
We introduce a lightweight and accurate localization method that only utilizes t he geometry of 2D-3D lines. Given a pre-captured 3D map our approach localizes a panorama image taking advantage of the holistic 360 degree view. The system mit igates potential privacy breaches or domain discrepancies by avoiding trained or hand-crafted visual descriptors. However as lines alone can be ambiguous we exp ress distinctive yet compact spatial contexts from relationships between lines n amely the dominant directions of parallel lines and the intersection between non -parallel lines. The resulting representations are efficient in processing time and memory compared to conventional visual descriptor-based methods. Given the g roups of dominant line directions and their intersections we accelerate the sear ch process to test thousands of pose candidates in less than a millisecond witho ut sacrificing accuracy. We empirically show that the proposed 2D-3D matching ca n localize panoramas for challenging scenes with similar structures dramatic dom ain shifts or illumination changes. Our fully geometric approach does not involv e extensive parameter tuning or neural network training making it a practical al gorithm that can be readily deployed in the real world. Project page including t he code is available through this link: https://82magnolia.github.io/fgpl/.
****
CAPE: CAM as a Probabilistic Ensemble for Enhanced DNN Interpretation
Townim Faisal Chowdhury, Kewen Liao, Vu Minh Hieu Phan, Minh-Son To, Yutong Xie, Kevin Hung, David Ross, Anton van den Hengel, Johan W. Verjans, Zhibin Liao; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11072-11081
Deep Neural Networks (DNNs) are widely used for visual classification tasks but their complex computation process and black-box nature hinder decision transpare ncy and interpretability. Class activation maps (CAMs) and recent variants provi de ways to visually explain the DNN decision-making process by displaying 'atten tion' heatmaps of the DNNs. Nevertheless the CAM explanation only offers relativ e attention information that is on an attention heatmap we can interpret which i mage region is more or less important than the others. However these regions can not be meaningfully compared across classes and the contribution of each region to the model's class prediction is not revealed. To address these challenges tha t ultimately lead to better DNN Interpretation in this paper we propose CAPE a n ovel reformulation of CAM that provides a unified and probabilistically meaningf ul assessment of the contributions of image regions. We quantitatively and quali tatively compare CAPE with state-of-the-art CAM methods on CUB and ImageNet benc hmark datasets to demonstrate enhanced interpretability. We also test on a cytol ogy imaging dataset depicting a challenging Chronic Myelomonocytic Leukemia (CMM L) diagnosis problem. Code is available at:https://github.com/AIML-MED/CAPE.
****
NeRF Director: Revisiting View Selection in Neural Volume Rendering
Wenhui Xiao, Rodrigo Santa Cruz, David Ahmedt-Aristizabal, Olivier Salvado, Clin ton Fookes, Leo Lebrat; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2024, pp. 20742-20751

Neural Rendering representations have significantly contributed to the field of
3D computer vision. Given their potential considerable efforts have been investe
d to improve their performance. Nonetheless the essential question of selecting
training views is yet to be thoroughly investigated. This key aspect plays a vit
al role in achieving high-quality results and aligns with the well-known tenet o
f deep learning: "garbage in garbage out". In this paper we first illustrate the
 importance of view selection by demonstrating how a simple rotation of the test
 views within the most pervasive NeRF dataset can lead to consequential shifts i
n the performance rankings of state-of-the-art techniques. To address this chall
enge we introduce a unified framework for view selection methods and devise a th
orough benchmark to assess its impact. Significant improvements can be achieved
without leveraging error or uncertainty estimation but focusing on uniform view
coverage of the reconstructed object resulting in a training-free approach. Usin
g this technique we show that high-quality renderings can be achieved faster by
using fewer views. We conduct extensive experiments on both synthetic datasets a
nd realistic data to demonstrate the effectiveness of our proposed method compar
ed with random conventional error-based and uncertainty-guided view selection.
****
Taming the Tail in Class-Conditional GANs: Knowledge Sharing via Unconditional T
raining at Lower Resolutions
Saeed Khorram, Mingqi Jiang, Mohamad Shahbazi, Mohamad H. Danesh, Li Fuxin; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 7580-7590
Despite extensive research on training generative adversarial networks (GANs) wi
th limited training data learning to generate images from long-tailed training d
istributions remains fairly unexplored. In the presence of imbalanced multi-clas
s training data GANs tend to favor classes with more samples leading to the gene
ration of low quality and less diverse samples in tail classes. In this study we
 aim to improve the training of class-conditional GANs with long-tailed data. We
 propose a straightforward yet effective method for knowledge sharing allowing t
ail classes to borrow from the rich information from classes with more abundant
training data. More concretely we propose modifications to existing class-condit
ional GAN architectures to ensure that the lower-resolution layers of the genera
tor are trained entirely unconditionally while reserving class-conditional gener
ation for the higher-resolution layers. Experiments on several long-tail benchma
rks and GAN architectures demonstrate a significant improvement over existing me
thods in both the diversity and fidelity of the generated images. The code is av
ailable at https://github.com/khorrams/utlo.
****
VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Cor
respondence
Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangji
e Wu, David Junhao Zhang, Mike Zheng Shou, Kevin Tang; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7621-
7630
Current diffusion-based video editing primarily focuses on structure-preserved e
diting by utilizing various dense correspondences to ensure temporal consistency
 and motion alignment. However these approaches are often ineffective when the t
arget edit involves a shape change. To embark on video editing with shape change
 we explore customized video subject swapping in this work where we aim to repla
ce the main subject in a source video with a target subject having a distinct id
entity and potentially different shape. In contrast to previous methods that rel
y on dense correspondences we introduce the VideoSwap framework that exploits se
mantic point correspondences inspired by our observation that only a small numbe
r of semantic points are necessary to align the subject's motion trajectory and
modify its shape. We also introduce various user-point interactions (e.g. removi
ng points and dragging points) to address various semantic point correspondence.
 Extensive experiments demonstrate state-of-the-art video subject swapping resul
ts across a variety of real-world videos.
****

SonicVisionLM: Playing Sound with Vision Language Models
Zhifeng Xie, Shengye Yu, Qile He, Mengtian Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26866-26875
There has been a growing interest in the task of generating sound for silent videos primarily because of its practicality in streamlining video post-production. However existing methods for video-sound generation attempt to directly create sound from visual representations which can be challenging due to the difficulty of aligning visual representations with audio representations. In this paper we present SonicVisionLM a novel framework aimed at generating a wide range of sound effects by leveraging vision-language models(VLMs). Instead of generating audio directly from video we use the capabilities of powerful VLMs. When provided with a silent video our approach first identifies events within the video using a VLM to suggest possible sounds that match the video content. This shift in approach transforms the challenging task of aligning image and audio into more well-studied sub-problems of aligning image-to-text and text-to-audio through the popular diffusion models. To improve the quality of audio recommendations with LLMs we have collected an extensive dataset that maps text descriptions to specific sound effects and developed a time-controlled audio adapter. Our approach surpasses current state-of-the-art methods for converting video to audio enhancing synchronization with the visuals and improving alignment between audio and video components. Project page: https://yusiissy.github.io/SonicVisionLM.github.io/
****
Multi-Space Alignments Towards Universal LiDAR Segmentation
Youquan Liu, Lingdong Kong, Xiaoyang Wu, Runnan Chen, Xin Li, Liang Pan, Ziwei Liu, Yuexin Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14648-14661
A unified and versatile LiDAR segmentation model with strong robustness and generalizability is desirable for safe autonomous driving perception. This work presents M3Net a one-of-a-kind framework for fulfilling multi-task multi-dataset multi-modality LiDAR segmentation in a universal manner using just a single set of parameters. To better exploit data volume and diversity we first combine large-scale driving datasets acquired by different types of sensors from diverse scenes and then conduct alignments in three spaces namely data feature and label spaces during the training. As a result M3Net is capable of taming heterogeneous data for training state-of-the-art LiDAR segmentation models. Extensive experiments on twelve LiDAR segmentation datasets verify our effectiveness. Notably using a shared set of parameters M3Net achieves 75.1% 83.1% and 72.4% mIoU scores respectively on the official benchmarks of SemanticKITTI nuScenes and Waymo Open.
****
DiffuScene: Denoising Diffusion Models for Generative Indoor Scene Synthesis
Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20507-20518
We present DiffuScene for indoor 3D scene synthesis based on a novel scene configuration denoising diffusion model. It generates 3D instance properties stored in an unordered object set and retrieves the most similar geometry for each object configuration which is characterized as a concatenation of different attributes including location size orientation semantics and geometry features. We introduce a diffusion network to synthesize a collection of 3D indoor objects by denoising a set of unordered object attributes. Unordered parametrization simplifies and eases the joint distribution approximation. The shape feature diffusion facilitates natural object placements including symmetries. Our method enables many downstream applications including scene completion scene arrangement and text-conditioned scene synthesis. Experiments on the 3D-FRONT dataset show that our method can synthesize more physically plausible and diverse indoor scenes than state-of-the-art methods. Extensive ablation studies verify the effectiveness of our design choice in scene diffusion models.
****
Hierarchical Histogram Threshold Segmentation - Auto-terminating High-detail Oversegmentation

Thomas V. Chang, Simon Seibt, Bartosz von Rymon Lipinski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3195-3204

Superpixels play a crucial role in image processing by partitioning an image into clusters of pixels with similar visual attributes. This facilitates subsequent image processing tasks offering computational advantages over the manipulation of individual pixels. While numerous oversegmentation techniques have emerged in recent years many rely on predefined initialization and termination criteria. In this paper a novel top-down superpixel segmentation algorithm called Hierarchical Histogram Threshold Segmentation (HHTS) is introduced. It eliminates the need for initialization and implements auto-termination outperforming state-of-the-art methods w.r.t boundary recall. This is achieved by iteratively partitioning individual pixel segments into foreground and background and applying intensity thresholding across multiple color channels. The underlying iterative process constructs a superpixel hierarchy that adapts to local detail distributions until color information exhaustion. Experimental results demonstrate the superiority of the proposed approach in terms of boundary adherence while maintaining competitive runtime performance on the BSDS500 and NYUV2 datasets. Furthermore an application of HHTS in refining machine learning-based semantic segmentation masks produced by the Segment Anything Foundation Model (SAM) is presented.

****

Once for Both: Single Stage of Importance and Sparsity Search for Vision Transformer Compression

Hancheng Ye, Chong Yu, Peng Ye, Renqiu Xia, Yansong Tang, Jiwen Lu, Tao Chen, Bo Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5578-5588

Recent Vision Transformer Compression (VTC) works mainly follow a two-stage scheme where the importance score of each model unit is first evaluated or preset in each submodule followed by the sparsity score evaluation according to the target sparsity constraint. Such a separate evaluation process induces the gap between importance and sparsity score distributions thus causing high search costs for VTC. In this work for the first time we investigate how to integrate the evaluations of importance and sparsity scores into a single stage searching the optimal subnets in an efficient manner. Specifically we present OFB a cost-efficient approach that simultaneously evaluates both importance and sparsity scores termed Once for Both (OFB) for VTC. First a bi-mask scheme is developed by entangling the importance score and the differentiable sparsity score to jointly determine the pruning potential (prunability) of each unit. Such a bi-mask search strategy is further used together with a proposed adaptive one-hot loss to realize the progressive-and-efficient search for the most important subnet. Finally Progressive Masked Image Modeling (PMIM) is proposed to regularize the feature space to be more representative during the search process which may be degraded by the dimension reduction. Extensive experiments demonstrate that OFB can achieve superior compression performance over state-of-the-art searching-based and pruning-based methods under various Vision Transformer architectures meanwhile promoting search efficiency significantly e.g. costing one GPU search day for the compression of DeiT-S on ImageNet-1K.

****

As-Plausible-As-Possible: Plausibility-Aware Mesh Deformation Using 2D Diffusion Priors

Seungwoo Yoo, Kunho Kim, Vladimir G. Kim, Minhyuk Sung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4315-4324

We present As-Plausible-as-Possible (APAP) mesh deformation technique that leverages 2D diffusion priors to preserve the plausibility of a mesh under user-controlled deformation. Our framework uses per-face Jacobians to represent mesh deformations where mesh vertex coordinates are computed via a differentiable Poisson Solve. The deformed mesh is rendered and the resulting 2D image is used in the Score Distillation Sampling (SDS) process which enables extracting meaningful plausibility priors from a pretrained 2D diffusion model. To better preserve the id

entity of the edited mesh we fine-tune our 2D diffusion model with LoRA. Gradients extracted by SDS and a user-prescribed handle displacement are then backpropagated to the per-face Jacobians and we use iterative gradient descent to compute the final deformation that balances between the user edit and the output plausibility. We evaluate our method with 2D and 3D meshes and demonstrate qualitative and quantitative improvements when using plausibility priors over geometry-preservation or distortion-minimization priors used by previous techniques. Our project page is at: https://as-plausible-aspossible.github.io/
****
MCNet: Rethinking the Core Ingredients for Accurate and Efficient Homography Estimation
Haokai Zhu, Si-Yuan Cao, Jianxin Hu, Sitong Zuo, Beinan Yu, Jiacheng Ying, Junwei Li, Hui-Liang Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25932-25941
We propose Multiscale Correlation searching homography estimation Network namely MCNet an iterative deep homography estimation architecture. Different from previous approaches that achieve iterative refinement by correlation searching within a single scale MCNet combines the multiscale strategy with correlation searching incurring nearly ignored computational overhead. Moreover MCNet adopts a Fine-Grained Optimization loss function named FGO loss to further boost the network training at the convergent stage which can improve the estimation accuracy without additional computational overhead. According to our experiments using the above two simple strategies can produce significant homography estimation accuracy with considerable efficiency. We show that MCNet achieves state-of-the-art performance on a variety of datasets including common scene MSCOCO cross-modal scene GoogleEarth and GoogleMap and dynamic scene SPID. Compared to the previous SOTA method 2-scale RHWF our MCNet reduces inference time FLOPs parameter cost and memory cost by 78.9% 73.5% 34.1% and 33.2% respectively while achieving 20.5% (MSCOCO) 43.4% (GoogleEarth) and 41.1% (GoogleMap) mean average corner error (MACE) reduction. Source code is available at https://github.com/zjuzhk/MCNet.
****
ECLIPSE: Efficient Continual Learning in Panoptic Segmentation with Visual Prompt Tuning
Beomyoung Kim, Joonsang Yu, Sung Ju Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3346-3356
Panoptic segmentation combining semantic and instance segmentation stands as a cutting-edge computer vision task. Despite recent progress with deep learning models the dynamic nature of real-world applications necessitates continual learning where models adapt to new classes (plasticity) over time without forgetting old ones (catastrophic forgetting). Current continual segmentation methods often rely on distillation strategies like knowledge distillation and pseudo-labeling which are effective but result in increased training complexity and computational overhead. In this paper we introduce a novel and efficient method for continual panoptic segmentation based on Visual Prompt Tuning dubbed ECLIPSE. Our approach involves freezing the base model parameters and fine-tuning only a small set of prompt embeddings addressing both catastrophic forgetting and plasticity and significantly reducing the trainable parameters. To mitigate inherent challenges such as error propagation and semantic drift in continual segmentation we propose logit manipulation to effectively leverage common knowledge across the classes. Experiments on ADE20K continual panoptic segmentation benchmark demonstrate the superiority of ECLIPSE notably its robustness against catastrophic forgetting and its reasonable plasticity achieving a new state-of-the-art. The code is available at https://github.com/clovaai/ECLIPSE.
****
Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters
Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, You He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23219-23230
Continual learning can empower vision-language models to continuously acquire ne

w knowledge without the need for access to the entire historical dataset. Howeve r mitigating the performance degradation in large-scale models is non-trivial du e to (i) parameter shifts throughout lifelong learning and (ii) significant comp utational burdens associated with full-model tuning. In this work we present a p arameter-efficient continual learning framework to alleviate long-term forgettin g in incremental learning with vision-language models. Our approach involves the dynamic expansion of a pre-trained CLIP model through the integration of Mixtur e-of-Experts (MoE) adapters in response to new tasks. To preserve the zero-shot recognition capability of vision-language models we further introduce a Distribu tion Discriminative Auto-Selector (DDAS) that automatically routes in-distributi on and out-of-distribution inputs to the MoE Adapter and the original CLIP respe ctively. Through extensive experiments across various settings our proposed meth od consistently outperforms previous state-of-the-art approaches while concurren tly reducing parameter training burdens by 60%.
****
MaGGIe: Masked Guided Gradual Human Instance Matting
Chuong Huynh, Seoung Wug Oh, Abhinav Shrivastava, Joon-Young Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024 , pp. 3870-3879
Human matting is a foundation task in image and video processing where human for eground pixels are extracted from the input. Prior works either improve the accu racy by additional guidance or improve the temporal consistency of a single inst ance across frames. We propose a new framework MaGGIe Masked Guided Gradual Huma n Instance Matting which predicts alpha mattes progressively for each human inst ances while maintaining the computational cost precision and consistency. Our me thod leverages modern architectures including transformer attention and sparse c onvolution to output all instance mattes simultaneously without exploding memory and latency. Although keeping constant inference costs in the multiple-instance scenario our framework achieves robust and versatile performance on our propose d synthesized benchmarks. With the higher quality image and video matting benchm arks the novel multi-instance synthesis approach from publicly available sources is introduced to increase the generalization of models in real-world scenarios. Our code and datasets are available at https://maggie-matt.github.io
****
FlowDiffuser: Advancing Optical Flow Estimation with Diffusion Models
Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, Shuaicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 024, pp. 19167-19176
Optical flow estimation a process of predicting pixel-wise displacement between consecutive frames has commonly been approached as a regression task in the age of deep learning. Despite notable advancements this de facto paradigm unfortunat ely falls short in generalization performance when trained on synthetic or const rained data. Pioneering a paradigm shift we reformulate optical flow estimation as a conditional flow generation challenge unveiling FlowDiffuser --- a new f amily of optical flow models that could have stronger learning and generalizatio n capabilities. FlowDiffuser estimates optical flow through a `noise-to-flow' st rategy progressively eliminating noise from randomly generated flows conditioned on the provided pairs. To optimize accuracy and efficiency our FlowDiffuser inc orporates a novel Conditional Recurrent Denoising Decoder (Conditional-RDD) stre amlining the flow estimation process. It incorporates a unique Hidden State Deno ising (HSD) paradigm effectively leveraging the information from previous time s teps. Moreover FlowDiffuser can be easily integrated into existing flow networks leading to significant improvements in performance metrics compared to conventi onal implementations. Experiments on challenging benchmarks including Sintel and KITTI demonstrate the effectiveness of our FlowDiffuser with superior performan ce to existing state-of-the-art models. Code is available at https://github.com/ LA30/FlowDiffuser.
****
Benchmarking Implicit Neural Representation and Geometric Rendering in Real-Time RGB-D SLAM

Tongyan Hua, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21346-21356

Implicit neural representation (INR) in combination with geometric rendering has recently been employed in real-time dense RGB-D SLAM. Despite active research endeavors being made there lacks a unified protocol for fair evaluation impeding the evolution of this area. In this work we establish to our knowledge the first open-source benchmark framework to evaluate the performance of a wide spectrum of commonly used INRs and rendering functions for mapping and localization. The goal of our benchmark is to 1) gain an intuition of how different INRs and rendering functions impact mapping and localization and 2) establish a unified evaluation protocol w.r.t. the design choices that may impact the mapping and localization. With the framework we conduct a large suite of experiments offering various insights in choosing the INRs and geometric rendering functions: for example the dense feature grid outperforms other INRs (e.g. tri-plane and hash grid) even when geometric and color features are jointly encoded for memory efficiency. To extend the findings into the practical scenario a hybrid encoding strategy is proposed to bring the best of the accuracy and completion from the grid-based and decomposition-based INRs. We further propose explicit hybrid encoding for high-fidelity dense grid mapping to comply with the RGB-D SLAM system that puts the premise on robustness and computation efficiency.

****

Free3D: Consistent Novel View Synthesis without 3D Representation

Chuanxia Zheng, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9720-9731

We introduce Free3D a simple accurate method for monocular open-set novel view synthesis (NVS). Similar to Zero-1-to-3 we start from a pre-trained 2D image generator for generalization and fine-tune it for NVS. Compared to other works that took a similar approach we obtain significant improvements without resorting to an explicit 3D representation which is slow and memory-consuming and without training an additional network for 3D reconstruction. Our key contribution is to improve the way the target camera pose is encoded in the network which we do by introducing a new ray conditioning normalization (RCN) layer. The latter injects pose information in the underlying 2D image generator by telling each pixel its viewing direction. We further improve multi-view consistency by using light-weight multi-view attention layers and by sharing generation noise between the different views. We train Free3D on the Objaverse dataset and demonstrate excellent generalization to new categories in new datasets including OmniObject3D and GSO. The project page is available at https://chuanxiaz.com/free3d/.

****

SuperSVG: Superpixel-based Scalable Vector Graphics Synthesis

Teng Hu, Ran Yi, Baihong Qian, Jiangning Zhang, Paul L. Rosin, Yu-Kun Lai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24892-24901

SVG (Scalable Vector Graphics) is a widely used graphics format that possesses excellent scalability and editability. Image vectorization that aims to convert raster images to SVGs is an important yet challenging problem in computer vision and graphics. Existing image vectorization methods either suffer from low reconstruction accuracy for complex images or require long computation time. To address this issue we propose SuperSVG a superpixel-based vectorization model that achieves fast and high-precision image vectorization. Specifically we decompose the input image into superpixels to help the model focus on areas with similar colors and textures. Then we propose a two-stage self-training framework where a coarse-stage model is employed to reconstruct the main structure and a refinement-stage model is used for enriching the details. Moreover we propose a novel dynamic path warping loss to help the refinement-stage model to inherit knowledge from the coarse-stage model. Extensive qualitative and quantitative experiments demonstrate the superior performance of our method in terms of reconstruction accuracy and inference time compared to state-of-the-art approaches. The code is available in https://github.com/sjtuplayer/SuperSVG.

****

AV2AV: Direct Audio-Visual Speech to Audio-Visual Speech Translation with Unifie
d Audio-Visual Speech Representation
Jeongsoo Choi, Se Jin Park, Minsu Kim, Yong Man Ro; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27325-27
337
This paper proposes a novel direct Audio-Visual Speech to Audio-Visual Speech Tr
anslation (AV2AV) framework where the input and output of the system are multimo
dal (i.e. audio and visual speech). With the proposed AV2AV two key advantages c
an be brought: 1) We can perform real-like conversations with individuals worldw
ide in a virtual meeting by utilizing our own primary languages. In contrast to
Speech-to-Speech Translation (A2A) which solely translates between audio modalit
ies the proposed AV2AV directly translates between audio-visual speech. This cap
ability enhances the dialogue experience by presenting synchronized lip movement
s along with the translated speech. 2) We can improve the robustness of the spok
en language translation system. By employing the complementary information of au
dio-visual speech the system can effectively translate spoken language even in t
he presence of acoustic noise showcasing robust performance. To mitigate the pro
blem of the absence of a parallel AV2AV translation dataset we propose to train
our spoken language translation system with the audio-only dataset of A2A. This
is done by learning unified audio-visual speech representations through self-sup
ervised learning in advance to train the translation system. Moreover we propose
 an AV-Renderer that can generate raw audio and video in parallel. It is designe
d with zero-shot speaker modeling thus the speaker in source audio-visual speech
 can be maintained at the target translated audio-visual speech. The effectivene
ss of AV2AV is evaluated with extensive experiments in a many-to-many language t
ranslation setting. Demo page is available on choijeongsoo.github.io/av2av.
****
Towards the Uncharted: Density-Descending Feature Perturbation for Semi-supervis
ed Semantic Segmentation
Xiaoyang Wang, Huihui Bai, Limin Yu, Yao Zhao, Jimin Xiao; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3
303-3312
Semi-supervised semantic segmentation allows model to mine effective supervision
 from unlabeled data to complement label-guided training. Recent research has pr
imarily focused on consistency regularization techniques exploring perturbation-
invariant training at both the image and feature levels. In this work we propose
d a novel feature-level consistency learning framework named Density-Descending
Feature Perturbation (DDFP). Inspired by the low-density separation assumption i
n semi-supervised learning our key insight is that feature density can shed a li
ght on the most promising direction for the segmentation classifier to explore w
hich is the regions with lower density. We propose to shift features with confid
ent predictions towards lower-density regions by perturbation injection. The per
turbed features are then supervised by the predictions on the original features
thereby compelling the classifier to explore less dense regions to effectively r
egularize the decision boundary. Central to our method is the estimation of feat
ure density. To this end we introduce a lightweight density estimator based on n
ormalizing flow allowing for efficient capture of the feature density distributi
on in an online manner. By extracting gradients from the density estimator we ca
n determine the direction towards less dense regions for each feature. The propo
sed DDFP outperforms other designs on feature-level perturbations and shows stat
e of the art performances on both Pascal VOC and Cityscapes dataset under variou
s partition protocols. The project is available at https://github.com/Gavinwxy/D
DFP.
****
WALT3D: Generating Realistic Training Data from Time-Lapse Imagery for Reconstru
cting Dynamic Objects Under Occlusion
Khiem Vuong, N Dinesh Reddy, Robert Tamburo, Srinivasa G. Narasimhan; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 9514-9524
Current methods for 2D and 3D object understanding struggle with severe occlusio

ns in busy urban environments partly due to the lack of large-scale labeled grou
nd-truth annotations for learning occlusion. In this work we introduce a novel f
ramework for automatically generating a large realistic dataset of dynamic objec
ts under occlusions using freely available time-lapse imagery. By leveraging off
-the-shelf 2D (bounding box segmentation keypoint) and 3D (pose shape) predictio
ns as pseudo-groundtruth unoccluded 3D objects are identified automatically and
composited into the background in a clip-art style ensuring realistic appearance
s and physically accurate occlusion configurations. The resulting clip-art image
 with pseudo-groundtruth enables efficient training of object reconstruction met
hods that are robust to occlusions. Our method demonstrates significant improvem
ents in both 2D and 3D reconstruction particularly in scenarios with heavily occ
luded objects like vehicles and people in urban scenes.
****
RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation
Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, Wenming Yang; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 1491-1500
Real-time multi-person pose estimation presents significant challenges in balanc
ing speed and precision. While two-stage top-down methods slow down as the numbe
r of people in the image increases existing one-stage methods often fail to simu
ltaneously deliver high accuracy and real-time performance. This paper introduce
s RTMO a one-stage pose estimation framework that seamlessly integrates coordina
te classification by representing keypoints using dual 1-D heatmaps within the Y
OLO architecture achieving accuracy comparable to top-down methods while maintai
ning high speed. We propose a dynamic coordinate classifier and a tailored loss
function for heatmap learning specifically designed to address the incompatibili
ties between coordinate classification and dense prediction models. RTMO outperf
orms state-of-the-art one-stage pose estimators achieving 1.1% higher AP on COCO
 while operating about 9 times faster with the same backbone. Our largest model
RTMO-l attains 74.8% AP on COCO val2017 and 141 FPS on a single V100 GPU demonst
rating its efficiency and accuracy. The code and models are available at https:/
/github.com/open-mmlab/mmpose/tree/main/projects/rtmo.
****
Contrastive Mean-Shift Learning for Generalized Category Discovery
Sua Choi, Dahyun Kang, Minsu Cho; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2024, pp. 23094-23104
We address the problem of generalized category discovery (GCD) that aims to part
ition a partially labeled collection of images; only a small part of the collect
ion is labeled and the total number of target classes is unknown. To address thi
s generalized image clustering problem we revisit the mean-shift algorithm i.e.
a classic powerful technique for mode seeking and incorporate it into a contrast
ive learning framework. The proposed method dubbed Contrastive Mean-Shift (CMS)
learning trains an embedding network to produce representations with better clus
tering properties by an iterative process of mean shift and contrastive update.
Experiments demonstrate that our method both in settings with and without the to
tal number of clusters being known achieves state-of-the-art performance on six
public GCD benchmarks without bells and whistles.
****
Towards Language-Driven Video Inpainting via Multimodal Large Language Models
Jianzong Wu, Xiangtai Li, Chenyang Si, Shangchen Zhou, Jingkang Yang, Jiangning
Zhang, Yining Li, Kai Chen, Yunhai Tong, Ziwei Liu, Chen Change Loy; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 12501-12511
We introduce a new task -- language-driven video inpainting which uses natural l
anguage instructions to guide the inpainting process. This approach overcomes th
e limitations of traditional video inpainting methods that depend on manually la
beled binary masks a process often tedious and labor-intensive. We present the R
emove Objects from Videos by Instructions (ROVI) dataset containing 5650 videos
and 9091 inpainting results to support training and evaluation for this task. We
 also propose a novel diffusion-based language-driven video inpainting framework

the first end-to-end baseline for this task integrating Multimodal Large Language Models to understand and execute complex language-based inpainting requests effectively. Our comprehensive results showcase the dataset's versatility and the model's effectiveness in various language-instructed inpainting scenarios. We have made datasets code and models publicly available at https://github.com/jianzongwu/Language-Driven-Video-Inpainting.
****
WaveFace: Authentic Face Restoration with Efficient Frequency Recovery
Yunqi Miao, Jiankang Deng, Jungong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6583-6592
Although diffusion models are rising as a powerful solution for blind face restoration they are criticized for two problems: 1) slow training and inference speed and 2) failure in preserving identity and recovering fine-grained facial details. In this work we propose WaveFace to solve the problems in the frequency domain where low- and high-frequency components decomposed by wavelet transformation are considered individually to maximize authenticity as well as efficiency. The diffusion model is applied to recover the low-frequency component only which presents general information of the original image but 1/16 in size. To preserve the original identity the generation is conditioned on the low-frequency component of low-quality images at each denoising step. Meanwhile high-frequency components at multiple decomposition levels are handled by a unified network which recovers complex facial details in a single step. Evaluations on four benchmark datasets show that: 1) WaveFace outperforms state-of-the-art methods in authenticity especially in terms of identity preservation and 2) authentic images are restored with the efficiency 10x faster than existing diffusion model-based BFR methods.
****
CLIP-KD: An Empirical Study of CLIP Model Distillation
Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, Yongjun Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15952-15962
Contrastive Language-Image Pre-training (CLIP) has become a promising language-supervised visual pre-training framework. This paper aims to distill small CLIP models supervised by a large teacher CLIP model. We propose several distillation strategies including relation feature gradient and contrastive paradigms to examine the effectiveness of CLIP-Knowledge Distillation (KD). We show that a simple feature mimicry with Mean Squared Error loss works surprisingly well. Moreover interactive contrastive learning across teacher and student encoders is also effective in performance improvement. We explain that the success of CLIP-KD can be attributed to maximizing the feature similarity between teacher and student. The unified method is applied to distill several student models trained on CC3M+12M. CLIP-KD improves student CLIP models consistently over zero-shot ImageNet classification and cross-modal retrieval benchmarks. When using ViT-L/14 pretrained on Laion-400M as the teacher CLIP-KD achieves 57.5% and 55.4% zero-shot top-1 ImageNet accuracy over ViT-B/16 and ResNet-50 surpassing the original CLIP without KD by 20.5% and 20.1% margins respectively. Our code is released on https://github.com/winycg/CLIP-KD.
****
UltrAvatar: A Realistic Animatable 3D Avatar Diffusion Model with Authenticity Guided Textures
Mingyuan Zhou, Rakib Hyder, Ziwei Xuan, Guojun Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1238-1248
Recent advances in 3D avatar generation have gained significant attention. These breakthroughs aim to produce more realistic animatable avatars narrowing the gap between virtual and real-world experiences. Most of existing works employ Score Distillation Sampling (SDS) loss combined with a differentiable renderer and text condition to guide a diffusion model in generating 3D avatars. However SDS often generates over-smoothed results with few facial details thereby lacking the diversity compared with ancestral sampling. On the other hand other works generate 3D avatar from a single image where the challenges of unwanted lighting effe

cts perspective views and inferior image quality make them difficult to reliably reconstruct the 3D face meshes with the aligned complete textures. In this paper we propose a novel 3D avatar generation approach termed UltrAvatar with enhanced fidelity of geometry and superior quality of physically based rendering (PBR) textures without unwanted lighting. To this end the proposed approach presents a diffuse color extraction model and an authenticity guided texture diffusion model. The former removes the unwanted lighting effects to reveal true diffuse colors so that the generated avatars can be rendered under various lighting conditions. The latter follows two gradient-based guidances for generating PBR textures to render diverse face-identity features and details better aligning with 3D mesh geometry. We demonstrate the effectiveness and robustness of the proposed method outperforming the state-of-the-art methods by a large margin in the experiments.
****
OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning
Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, Wenqiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19079-19091
Visual object tracking aims to localize the target object of each frame based on its initial appearance in the first frame. Depending on the input modility tracking tasks can be divided into RGB tracking and RGB+X (e.g. RGB+N and RGB+D) tracking. Despite the different input modalities the core aspect of tracking is the temporal matching. Based on this common ground we present a general framework to unify various tracking tasks termed as OneTracker. OneTracker first performs a large-scale pre-training on a RGB tracker called Foundation Tracker. This pretraining phase equips the Foundation Tracker with a stable ability to estimate the location of the target object. Then we regard other modality information as prompt and build Prompt Tracker upon Foundation Tracker. Through freezing the Foundation Tracker and only adjusting some additional trainable parameters Prompt Tracker inhibits the strong localization ability from Foundation Tracker and achieves parameter-efficient finetuning on downstream RGB+X tracking tasks. To evaluate the effectiveness of our general framework OneTracker which is consisted of Foundation Tracker and Prompt Tracker we conduct extensive experiments on 6 popular tracking tasks across 11 benchmarks and our OneTracker outperforms other models and achieves state-of-the-art performance.
****
SC-Tune: Unleashing Self-Consistent Referential Comprehension in Large Vision Language Models
Tongtian Yue, Jie Cheng, Longteng Guo, Xingyuan Dai, Zijia Zhao, Xingjian He, Gang Xiong, Yisheng Lv, Jing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13073-13083
Recent trends in Large Vision Language Models (LVLMs) research have been increasingly focusing on advancing beyond general image understanding towards more nuanced object-level referential comprehension. In this paper we present and delve into the self-consistency capability of LVLMs a crucial aspect that reflects the models' ability to both generate informative captions for specific objects and subsequently utilize these captions to accurately re-identify the objects in a closed-loop process. This capability significantly mirrors the precision and reliability of fine-grained visual-language understanding. Our findings reveal that the self-consistency level of existing LVLMs falls short of expectations posing limitations on their practical applicability and potential. To address this gap we introduce a novel fine-tuning paradigm named Self-Consistency Tuning (SC-Tune). It features the synergistic learning of a cyclic describer-locator system. This paradigm is not only data-efficient but also exhibits generalizability across multiple LVLMs. Through extensive experiments we demonstrate that SC-Tune significantly elevates performance across a spectrum of object-level vision-language benchmarks and maintains competitive or improved performance on image-level vision-language benchmarks. Both our model and code will be publicly available at htt

ps://github.com/ivattyue/SC-Tune.
****
Improving Depth Completion via Depth Feature Upsampling
Yufei Wang, Ge Zhang, Shaoqian Wang, Bo Li, Qi Liu, Le Hui, Yuchao Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21104-21113
The encoder-decoder network (ED-Net) is a commonly employed choice for existing depth completion methods but its working mechanism is ambiguous. In this paper we visualize the internal feature maps to analyze how the network densifies the input sparse depth. We find that the encoder feature of ED-Net focus on the areas with input depth points around. To obtain a dense feature and thus estimate complete depth the decoder feature tends to complement and enhance the encoder feature by skip-connection to make the fused encoder-decoder feature dense resulting in the decoder feature also exhibits sparse. However ED-Net obtains the sparse decoder feature from the dense fused feature at the previous stage where the "dense to sparse" process destroys the completeness of features and loses information. To address this issue we present a depth feature upsampling network (DFU) that explicitly utilizes these dense features to guide the upsampling of a low-resolution (LR) depth feature to a high-resolution (HR) one. The completeness of features is maintained throughout the upsampling process thus avoiding information loss. Furthermore we propose a confidence-aware guidance module (CGM) which is confidence-aware and performs guidance with adaptive receptive fields (GARF) to fully exploit the potential of these dense features as guidance. Experimental results show that our DFU a plug-and-play module can significantly improve the performance of existing ED-Net based methods with limited computational overheads and new SOTA results are achieved. Besides the generalization capability on sparser depth is also enhanced. Project page: https://npucvr.github.io/DFU.
****
NeRSP: Neural 3D Reconstruction for Reflective Objects with Sparse Polarized Images
Yufei Han, Heng Guo, Koki Fukai, Hiroaki Santo, Boxin Shi, Fumio Okura, Zhanyu Ma, Yunpeng Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11821-11830
We present NeRSP a Neural 3D reconstruction technique for Reflective surfaces with Sparse Polarized images. Reflective surface reconstruction is extremely challenging as specular reflections are view-dependent and thus violate the multiview consistency for multiview stereo. On the other hand sparse image inputs as a practical capture setting commonly cause incomplete or distorted results due to the lack of correspondence matching. This paper jointly handles the challenges from sparse inputs and reflective surfaces by leveraging polarized images. We derive photometric and geometric cues from the polarimetric image formation model and multiview azimuth consistency which jointly optimize the surface geometry modeled via implicit neural representation. Based on the experiments on our synthetic and real datasets we achieve the state-of-the-art surface reconstruction results with only 6 views as input.
****
Retrieval-Augmented Embodied Agents
Yichen Zhu, Zhicai Ou, Xiaofeng Mou, Jian Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17985-17995
Embodied agents operating in complex and uncertain environments face considerable challenges. While some advanced agents handle complex manipulation tasks with proficiency their success often hinges on extensive training data to develop their capabilities. In contrast humans typically rely on recalling past experiences and analogous situations to solve new problems. Aiming to emulate this human approach in robotics we introduce the Retrieval-Augmented Embodied Agent (RAEA). This innovative system equips robots with a form of shared memory significantly enhancing their performance. Our approach integrates a policy retriever allowing robots to access relevant strategies from an external policy memory bank based on multi-modal inputs. Additionally a policy generator is employed to assimilate these strategies into the learning process enabling robots to formulate effectiv

e responses to tasks. Extensive testing of RAEA in both simulated and real-world scenarios demonstrates its superior performance over traditional methods representing a major leap forward in robotic technology.
****
SAFDNet: A Simple and Effective Network for Fully Sparse 3D Object Detection
Gang Zhang, Junnan Chen, Guohuan Gao, Jianmin Li, Si Liu, Xiaolin Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14477-14486
LiDAR-based 3D object detection plays an essential role in autonomous driving. Existing high-performing 3D object detectors usually build dense feature maps in the backbone network and prediction head. However the computational costs introduced by the dense feature maps grow quadratically as the perception range increases making these models hard to scale up to long-range detection. Some recent works have attempted to construct fully sparse detectors to solve this issue; nevertheless the resulting models either rely on a complex multi-stage pipeline or exhibit inferior performance. In this work we propose a fully sparse adaptive feature diffusion network (SAFDNet) for LiDAR-based 3D object detection. In SAFDNet an adaptive feature diffusion strategy is designed to address the center feature missing problem. We conducted extensive experiments on Waymo Open nuScenes and Argoverse2 datasets. SAFDNet performed slightly better than the previous SOTA on the first two datasets but much better on the last dataset which features long-range detection verifying the efficacy of SAFDNet in scenarios where long-range detection is required. Notably on Argoverse2 SAFDNet surpassed the previous best hybrid detector HEDNet by 2.6% mAP while being 2.1x faster and yielded 2.1% mAP gains over the previous best sparse detector FSDv2 while being 1.3x faster. The code will be available at https://github.com/zhanggang001/HEDNet.
****
Attention-Propagation Network for Egocentric Heatmap to 3D Pose Lifting
Taeho Kang, Youngki Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 842-851
We present EgoTAP a heatmap-to-3D pose lifting method for highly accurate stereo egocentric 3D pose estimation. Severe self-occlusion and out-of-view limbs in egocentric camera views make accurate pose estimation a challenging problem. To address the challenge prior methods employ joint heatmaps-probabilistic 2D representations of the body pose but heatmap-to-3D pose conversion still remains an inaccurate process. We propose a novel heatmap-to-3D lifting method composed of the Grid ViT Encoder and the Propagation Network. The Grid ViT Encoder summarizes joint heatmaps into effective feature embedding using self-attention. Then the Propagation Network estimates the 3D pose by utilizing skeletal information to better estimate the position of obscure joints. Our method significantly outperforms the previous state-of-the-art qualitatively and quantitatively demonstrated by a 23.9% reduction of error in an MPJPE metric. Our source code is available on GitHub.
****
OmniMotionGPT: Animal Motion Generation with Limited Data
Zhangsihao Yang, Mingyuan Zhou, Mengyi Shan, Bingbing Wen, Ziwei Xuan, Mitch Hill, Junjie Bai, Guo-Jun Qi, Yalin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1249-1259
Our paper aims to generate diverse and realistic animal motion sequences from textual descriptions without a large-scale animal text-motion dataset. While the task of text-driven human motion synthesis is already extensively studied and benchmarked it remains challenging to transfer this success to other skeleton structures with limited data. In this work we design a model architecture that imitates Generative Pretraining Transformer (GPT) utilizing prior knowledge learned from human data to the animal domain. We jointly train motion autoencoders for both animal and human motions and at the same time optimize through the similarity scores among human motion encoding animal motion encoding and text CLIP embedding. Presenting the first solution to this problem we are able to generate animal motions with high diversity and fidelity quantitatively and qualitatively outperforming the results of training human motion generation baselines on animal data

. Additionally we introduce AnimalML3D the first text-animal motion dataset with 1240 animation sequences spanning 36 different animal identities. We hope this dataset would mediate the data scarcity problem in text-driven animal motion generation providing a new playground for the research community.
****
SNI-SLAM: Semantic Neural Implicit SLAM
Siting Zhu, Guangming Wang, Hermann Blum, Jiuming Liu, Liang Song, Marc Pollefeys, Hesheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21167-21177
We propose SNI-SLAM a semantic SLAM system utilizing neural implicit representation that simultaneously performs accurate semantic mapping high-quality surface reconstruction and robust camera tracking. In this system we introduce hierarchical semantic representation to allow multi-level semantic comprehension for top-down structured semantic mapping of the scene. In addition to fully utilize the correlation between multiple attributes of the environment we integrate appearance geometry and semantic features through cross-attention for feature collaboration. This strategy enables a more multifaceted understanding of the environment thereby allowing SNI-SLAM to remain robust even when single attribute is defective. Then we design an internal fusion-based decoder to obtain semantic RGB Truncated Signed Distance Field (TSDF) values from multi-level features for accurate decoding. Furthermore we propose a feature loss to update the scene representation at the feature level. Compared with low-level losses such as RGB loss and depth loss our feature loss is capable of guiding the network optimization on a higher-level. Our SNI-SLAM method demonstrates superior performance over all recent NeRF-based SLAM methods in terms of mapping and tracking accuracy on Replica and ScanNet datasets while also showing excellent capabilities in accurate semantic segmentation and real-time semantic mapping. Codes will be available at https://github.com/IRMVLab/SNI-SLAM.
****
InstanceDiffusion: Instance-level Control for Image Generation
Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, Ishan Misra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6232-6242
Text-to-image diffusion models produce high quality images but do not offer control over individual instances in the image. We introduce InstanceDiffusion that adds precise instance-level control to text-to-image diffusion models. InstanceDiffusion supports free-form language conditions per instance and allows flexible ways to specify instance locations such as simple single points scribbles bounding boxes or intricate instance segmentation masks and combinations thereof. We propose three major changes to text-to-image models that enable precise instance-level control. Our UniFusion block enables instance-level conditions for text-to-image models the ScaleU block improves image fidelity and our Multi-instance Sampler improves generations for multiple instances. InstanceDiffusion significantly surpasses specialized state-of-the-art models for each location condition. Notably on the COCO dataset we outperform previous state-of-the-art by 20.4% AP50box for box inputs and 25.4% IoU for mask inputs.
****
Unifying Top-down and Bottom-up Scanpath Prediction Using Transformers
Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, Dimitris Samaras; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1683-1693
Most models of visual attention aim at predicting either top-down or bottom-up control as studied using different visual search and free-viewing tasks. In this paper we propose the Human Attention Transformer (HAT) a single model that predicts both forms of attention control. HAT uses a novel transformer-based architecture and a simplified foveated retina that collectively create a spatio-temporal awareness akin to the dynamic visual working memory of humans. HAT not only establishes a new state-of-the-art in predicting the scanpath of fixations made during target-present and target-absent visual search and "taskless" free viewing but also makes human gaze behavior interpretable. Unlike previous methods that re

ly on a coarse grid of fixation cells and experience information loss due to fix
ation discretization HAT features a sequential dense prediction architecture and
 outputs a dense heatmap for each fixation thus avoiding discretizing fixations.
 HAT sets a new standard in computational attention which emphasizes effectivene
ss generality and interpretability. HAT's demonstrated scope and applicability w
ill likely inspire the development of new attention models that can better predi
ct human behavior in various attention-demanding scenarios. Code is available at
 https://github.com/cvlab-stonybrook/HAT.
****
HINTED: Hard Instance Enhanced Detector with Mixed-Density Feature Fusion for Sp
arsely-Supervised 3D Object Detection
Qiming Xia, Wei Ye, Hai Wu, Shijia Zhao, Leyuan Xing, Xun Huang, Jinhao Deng, Xi
n Li, Chenglu Wen, Cheng Wang; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2024, pp. 15321-15330
Current sparsely-supervised object detection methods largely depend on high thre
shold settings to derive high-quality pseudo labels from detector predictions. H
owever hard instances within point clouds frequently display incomplete structur
es causing decreased confidence scores in their assigned pseudo-labels. Previous
 methods inevitably result in inadequate positive supervision for these instance
s. To address this problem we propose a novel Hard INsTance Enhanced Detector HI
NTED for sparsely-supervised 3D object detection. Firstly we design a self-boost
ing teacher SBT model to generate more potential pseudo-labels enhancing the eff
ectiveness of information transfer. Then we introduce a mixed-density student MD
S model to concentrate on hard instances during the training phase thereby impro
ving detection accuracy. Our extensive experiments on the KITTI dataset validate
 our method's superior performance. Compared with leading sparsely-supervised me
thods HINTED significantly improves the detection performance on hard instances
notably outperforming fully-supervised methods in detecting challenging categori
es like cyclists. HINTED also significantly outperforms the state-of-the-art sem
i-supervised method on challenging categories. The code is available at https://
github.com/xmuqimingxia/HINTED.
****
Structured Gradient-based Interpretations via Norm-Regularized Adversarial Train
ing
Shizhan Gong, Qi Dou, Farzan Farnia; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2024, pp. 11009-11018
Gradient-based saliency maps have been widely used to explain the decisions of d
eep neural network classifiers. However standard gradient-based interpretation m
aps including the simple gradient and integrated gradient algorithms often lack
desired structures such as sparsity and connectedness in their application to re
al-world computer vision models. A common approach to induce sparsity-based stru
ctures into gradient-based saliency maps is to modify the simple gradient scheme
 using sparsification or norm-based regularization. However one drawback with su
ch post-processing approaches is the potentially significant loss in fidelity to
 the original simple gradient map. In this work we propose to apply adversarial
training as an in-processing scheme to train neural networks with structured sim
ple gradient maps. We demonstrate an existing duality between the regularized no
rms of the adversarial perturbations and gradient-based maps whereby we design a
dversarial training schemes promoting sparsity and group-sparsity properties in
simple gradient maps. We present comprehensive numerical results to show the inf
luence of our proposed norm-based adversarial training methods on the standard g
radient-based maps of standard neural network architectures on benchmark image d
atasets.
****
Building a Strong Pre-Training Baseline for Universal 3D Large-Scale Perception
Haoming Chen, Zhizhong Zhang, Yanyun Qu, Ruixin Zhang, Xin Tan, Yuan Xie; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2024, pp. 19925-19935
An effective pre-training framework with universal 3D representations is extreme
ly desired in perceiving large-scale dynamic scenes. However establishing such a

n ideal framework that is both task-generic and label-efficient poses a challeng
e in unifying the representation of the same primitive across diverse scenes. Th
e current contrastive 3D pre-training methods typically follow a frame-level con
sistency which focuses on the 2D-3D relationships in each detached image. Such i
nconsiderate consistency greatly hampers the promising path of reaching an unive
rsal pre-training framework: (1) The cross-scene semantic self-conflict \textit
 i.e. the intense collision between primitive segments of the same semantics fr
om different scenes; (2) Lacking a globally unified bond that pushes the cross-s
cene semantic consistency into 3D representation learning. To address above chal
lenges we propose a CSC framework that puts a scene-level semantic consistency i
n the heart bridging the connection of the similar semantic segments across vari
ous scenes. To achieve this goal we combine the coherent semantic cues provided
by the vision foundation model and the knowledge-rich cross-scene prototypes der
ived from the complementary multi-modality information. These allow us to train
a universal 3D pre-training model that facilitates various downstream tasks with
 less fine-tuning efforts. Empirically we achieve consistent improvements over S
OTA pre-training approaches in semantic segmentation (+1.4% mIoU) object detecti
on (+1.0% mAP) and panoptic segmentation (+3.0% PQ) using their task-specific 3D
 network on nuScenes. Code is released at \href https://github.com/chenhaomingbo
b/CSC https://github.com/chenhaomingbob/CSC hoping to inspire future research.
****
DS-NeRV: Implicit Neural Video Representation with Decomposed Static and Dynamic
 Codes
Hao Yan, Zhihui Ke, Xiaobo Zhou, Tie Qiu, Xidong Shi, Dadong Jiang; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 23019-23029
Implicit neural representations for video (NeRV) have recently become a novel wa
y for high-quality video representation. However existing works employ a single
network to represent the entire video which implicitly confuse static and dynami
c information. This leads to an inability to effectively compress the redundant
static information and lack the explicitly modeling of global temporal-coherent
dynamic details. To solve above problems we propose DS-NeRV which decomposes vid
eos into sparse learnable static codes and dynamic codes without the need for ex
plicit optical flow or residual supervision. By setting different sampling rates
 for two codes and applying weighted sum and interpolation sampling methods DS-N
eRV efficiently utilizes redundant static information while maintaining high-fre
quency details. Additionally we design a cross-channel attention-based (CCA) fus
ion module to efficiently fuse these two codes for frame decoding. Our approach
achieves a high quality reconstruction of 31.2 PSNR with only 0.35M parameters t
hanks to separate static and dynamic codes representation and outperforms existi
ng NeRV methods in many downstream tasks. Our project website is at https://haoy
an14.github.io/DS-NeRV.
****
3D-Aware Face Editing via Warping-Guided Latent Direction Learning
Yuhao Cheng, Zhuo Chen, Xingyu Ren, Wenhan Zhu, Zhengqin Xu, Di Xu, Changpeng Ya
ng, Yichao Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2024, pp. 916-926
3D facial editing a longstanding task in computer vision with broad applications
 is expected to fast and intuitively manipulate any face from arbitrary viewpoin
ts following the user's will. Existing works have limitations in terms of intuit
iveness generalization and efficiency. To overcome these challenges we propose F
aceEdit3D which allows users to directly manipulate 3D points to edit a 3D face
achieving natural and rapid face editing. After one or several points are manipu
lated by users we propose the tri-plane warping to directly manipulate the view-
independent 3D representation. To address the problem of distortion caused by tr
i-plane warping we train a warp-aware encoder to project the warped face onto a
standardized latent space. In this space we further propose directional latent e
diting to mitigate the identity bias caused by the encoder and realize the disen
tangled editing of various attributes. Extensive experiments show that our metho
d achieves superior results with rich facial details and nice identity preservat

ion. Our approach also supports general applications like multi-attribute contin
uous editing and cat/car editing. The project website is https://cyh-sj.github.i
o/FaceEdit3D/.
****
3DFIRES: Few Image 3D REconstruction for Scenes with Hidden Surfaces
Linyi Jin, Nilesh Kulkarni, David F. Fouhey; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9742-9751
This paper introduces 3DFIRES a novel system for scene-level 3D reconstruction f
rom posed images. Designed to work with as few as one view 3DFIRES reconstructs
the complete geometry of unseen scenes including hidden surfaces. With multiple
view inputs our method produces full reconstruction within all camera frustums.
A key feature of our approach is the fusion of multi-view information at the fea
ture level enabling the production of coherent and comprehensive 3D reconstructi
on. We train our system on non-watertight scans from large-scale real scene data
set. We show it matches the efficacy of single-view reconstruction methods with
only one input and surpasses existing techniques in both quantitative and qualit
ative measures for sparse-view 3D reconstruction.
****
CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation
Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, Seung
ryong Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2024, pp. 4113-4123
Open-vocabulary semantic segmentation presents the challenge of labeling each pi
xel within an image based on a wide range of text descriptions. In this work we
introduce a novel cost-based approach to adapt vision-language foundation models
 notably CLIP for the intricate task of semantic segmentation. Through aggregati
ng the cosine similarity score i.e. the cost volume between image and text embed
dings our method potently adapts CLIP for segmenting seen and unseen classes by
fine-tuning its encoders addressing the challenges faced by existing methods in
handling unseen classes. Building upon this we explore methods to effectively ag
gregate the cost volume considering its multi-modal nature of being established
between image and text embeddings. Furthermore we examine various methods for ef
ficiently fine-tuning CLIP.
****
Focus on Your Instruction: Fine-grained and Multi-instruction Image Editing by A
ttention Modulation
Qin Guo, Tianwei Lin; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2024, pp. 6986-6996
Recently diffusion-based methods like InstructPix2Pix (IP2P) have achieved effec
tive instruction-based image editing requiring only natural language instruction
s from the user. However these methods often inadvertently alter unintended area
s and struggle with multi-instruction editing resulting in compromised outcomes.
 To address these issues we introduce the Focus on Your Instruction (FoI) a meth
od designed to ensure precise and harmonious editing across multiple instruction
s without extra training or test-time optimization. In the FoI we primarily emph
asize two aspects: (1) precisely extracting regions of interest for each instruc
tion and (2) guiding the denoising process to concentrate within these regions o
f interest. For the first objective we identify the implicit grounding capabilit
y of IP2P from the cross-attention between instruction and image then develop an
 effective mask extraction method.
****
SDSTrack: Self-Distillation Symmetric Adapter Learning for Multi-Modal Visual Ob
ject Tracking
Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai T
ang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, Yong Liu; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 265
51-26561
Multimodal Visual Object Tracking (VOT) has recently gained significant attentio
n due to its robustness. Early research focused on fully fine-tuning RGB-based t
rackers which was inefficient and lacked generalized representation due to the s

carcity of multimodal data. Therefore recent studies have utilized prompt tuning to transfer pre-trained RGB-based trackers to multimodal data. However the modality gap limits pre-trained knowledge recall and the dominance of the RGB modality persists preventing the full utilization of information from other modalities. To address these issues we propose a novel symmetric multimodal tracking framework called SDSTrack. We introduce lightweight adaptation for efficient fine-tuning which directly transfers the feature extraction ability from RGB to other domains with a small number of trainable parameters and integrates multimodal features in a balanced symmetric manner. Furthermore we design a complementary masked patch distillation strategy to enhance the robustness of trackers in complex environments such as extreme weather poor imaging and sensor failure. Extensive experiments demonstrate that SDSTrack outperforms state-of-the-art methods in various multimodal tracking scenarios including RGB+Depth RGB+Thermal and RGB+Event tracking and exhibits impressive results in extreme conditions. Our source code is available at : https://github.com/hoqolo/SDSTrack.
****
MCPNet: An Interpretable Classifier via Multi-Level Concept Prototypes
Bor-Shiun Wang, Chien-Yi Wang, Wei-Chen Chiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10885-10894
Recent advancements in post-hoc and inherently interpretable methods have markedly enhanced the explanations of black box classifier models. These methods operate either through post-analysis or by integrating concept learning during model training. Although being effective in bridging the semantic gap between a model's latent space and human interpretation these explanation methods only partially reveal the model's decision-making process. The outcome is typically limited to high-level semantics derived from the last feature map. We argue that the explanations lacking insights into the decision processes at low and mid-level features are neither fully faithful nor useful. Addressing this gap we introduce the Multi-Level Concept Prototypes Classifier (MCPNet) an inherently interpretable model. MCPNet autonomously learns meaningful concept prototypes across multiple feature map levels using Centered Kernel Alignment (CKA) loss and an energy-based weighted PCA mechanism and it does so without reliance on predefined concept labels. Further we propose a novel classifier paradigm that learns and aligns multi-level concept prototype distributions for classification purposes via Class-aware Concept Distribution (CCD) loss. Our experiments reveal that our proposed MCPNet while being adaptable to various model architectures offers comprehensive multi-level explanations while maintaining classification accuracy. Additionally its concept distribution-based classification approach shows improved generalization capabilities in few-shot classification scenarios.
****
Semantic Shield: Defending Vision-Language Models Against Backdooring and Poisoning via Fine-grained Knowledge Alignment
Alvi Md Ishmam, Christopher Thomas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24820-24830
In recent years there has been enormous interest in vision-language models trained using self-supervised objectives. However the use of large-scale datasets scraped from the web for training also makes these models vulnerable to potential security threats such as backdooring and poisoning attacks. In this paper we propose a method for mitigating such attacks on contrastively trained vision-language models. Our approach leverages external knowledge extracted from a language model to prevent models from learning correlations between image regions which lack strong alignment with external knowledge. We do this by imposing constraints to enforce that attention paid by the model to visual regions is proportional to the alignment of those regions with external knowledge. We conduct extensive experiments using a variety of recent backdooring and poisoning attacks on multiple datasets and architectures. Our results clearly demonstrate that our proposed approach is highly effective at defending against such attacks across multiple settings while maintaining model utility and without requiring any changes at inference time.
****

AvatarGPT: All-in-One Framework for Motion Understanding Planning Generation and Beyond
Zixiang Zhou, Yu Wan, Baoyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1357-1366
Large Language Models(LLMs) have shown remarkable emergent abilities in unifying almost all (if not every) NLP tasks. In the human motion-related realm however researchers still develop siloed models for each task. Inspired by InstuctGPT[??] and the generalist concept behind Gato [??] we introduce AvatarGPT an All-in-One framework for motion understanding planning generations as well as other tasks such as motion in-between synthesis. AvatarGPT treats each task as one type of instruction fine-tuned on the shared LLM. All the tasks are seamlessly interconnected with language as the universal interface constituting a closed-loop within the framework. To achieve this human motion sequences are first encoded as discrete tokens which serve as the extended vocabulary of LLM. Then an unsupervised pipeline to generate natural language descriptions of human action sequences from in-the-wild videos is developed. Finally all tasks are jointly trained. Extensive experiments show that AvatarGPT achieves SOTA on low-level tasks and promising results on high-level tasks demonstrating the effectiveness of our proposed All-in-One framework. Moreover for the first time AvatarGPT enables a principled approach by iterative traversal of the tasks within the closed-loop for unlimited long-motion synthesis.
****
Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection
Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, Yunchao Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28130-28139
Recently the proliferation of highly realistic synthetic images facilitated through a variety of GANs and Diffusions has significantly heightened the susceptibility to misuse. While the primary focus of deepfake detection has traditionally centered on the design of detection algorithms an investigative inquiry into the generator architectures has remained conspicuously absent in recent years. This paper contributes to this lacuna by rethinking the architectures of CNN-based generator thereby establishing a generalized representation of synthetic artifacts. Our findings illuminate that the up-sampling operator can beyond frequency-based artifacts produce generalized forgery artifacts. In particular the local interdependence among image pixels caused by upsampling operators is significantly demonstrated in synthetic images generated by GAN or diffusion. Building upon this observation we introduce the concept of Neighboring Pixel Relationships(NPR) as a means to capture and characterize the generalized structural artifacts stemming from up-sampling operations. A comprehensive analysis is conducted on an open-world dataset comprising samples generated by 28 distinct generative models. This analysis culminates in the establishment of a novel state-of-the-art performance showcasing a remarkable 12.8% improvement over existing methods. The code is available at https://github.com/chuangchuangtan/NPR-DeepfakeDetection.
****
Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model
Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, Xiaofei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2263-2273
Co-speech gestures if presented in the lively form of videos can achieve superior visual effects in human-machine interaction. While previous works mostly generate structural human skeletons resulting in the omission of appearance information we focus on the direct generation of audio-driven co-speech gesture videos in this work. There are two main challenges: 1) A suitable motion feature is needed to describe complex human movements with crucial appearance information. 2) Gestures and speech exhibit inherent dependencies and should be temporally aligned even of arbitrary length. To solve these problems we present a novel motion-decoupled framework to generate co-speech gesture videos. Specifically we first introduce a well-designed nonlinear TPS transformation to obtain latent motion feat

ures preserving essential appearance information. Then a transformer-based diffusion model is proposed to learn the temporal correlation between gestures and speech and performs generation in the latent motion space followed by an optimal motion selection module to produce long-term coherent and consistent gesture videos. For better visual perception we further design a refinement network focusing on missing details of certain areas. Extensive experimental results show that our proposed framework significantly outperforms existing approaches in both motion and video-related evaluations. Our code demos and more resources are available at https://github.com/thuhcsi/S2G-MDDiffusion.
****
CDFormer: When Degradation Prediction Embraces Diffusion Model for Blind Image Super-Resolution
Qingguo Liu, Chenyi Zhuang, Pan Gao, Jie Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7455-7464
Existing Blind image Super-Resolution (BSR) methods focus on estimating either kernel or degradation information but have long overlooked the essential content details. In this paper we propose a novel BSR approach Content-aware Degradation-driven Transformer (CDFormer) to capture both degradation and content representations. However low-resolution images cannot provide enough content details and thus we introduce a diffusion-based module CDFormer_ diff to first learn Content Degradation Prior (CDP) in both low- and high-resolution images and then approximate the real distribution given only low-resolution information. Moreover we apply an adaptive SR network CDFormer_ SR that effectively utilizes CDP to refine features. Compared to previous diffusion-based SR methods we treat the diffusion model as an estimator that can overcome the limitations of expensive sampling time and excessive diversity. Experiments show that CDFormer can outperform existing methods establishing a new state-of-the-art performance on various benchmarks under blind settings. Codes and models will be available at https://github.com/I2-Multimedia-Lab/CDFormer.
****
HumanRef: Single Image to 3D Human Generation via Reference-Guided Diffusion
Jingbo Zhang, Xiaoyu Li, Qi Zhang, Yanpei Cao, Ying Shan, Jing Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1844-1854
Generating a 3D human model from a single reference image is challenging because it requires inferring textures and geometries in invisible views while maintaining consistency with the reference image. Previous methods utilizing 3D generative models are limited by the availability of 3D training data. Optimization-based methods that lift text-to-image diffusion models to 3D generation often fail to preserve the texture details of the reference image resulting in inconsistent appearances in different views. In this paper we propose HumanRef a 3D human generation framework from a single-view input. To ensure the generated 3D model is photorealistic and consistent with the input image HumanRef introduces a novel method called reference-guided score distillation sampling (Ref-SDS) which effectively incorporates image guidance into the generation process. Furthermore we introduce region-aware attention to Ref-SDS ensuring accurate correspondence between different body regions. Experimental results demonstrate that HumanRef outperforms state-of-the-art methods in generating 3D clothed humans with fine geometry photorealistic textures and view-consistent appearances. Code and model are available at https://eckertzhang.github.io/HumanRef.github.io/.
****
GlitchBench: Can Large Multimodal Models Detect Video Game Glitches?
Mohammad Reza Taesiri, Tianjun Feng, Cor-Paul Bezemer, Anh Nguyen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22444-22455
Large multimodal models (LMMs) have evolved from large language models (LLMs) to integrate multiple input modalities such as visual inputs. This integration augments the capacity of LLMs for tasks requiring visual comprehension and reasoning. However the extent and limitations of their enhanced abilities are not fully understood especially when it comes to real-world tasks. To address this gap we

introduce GlitchBench a novel benchmark derived from video game quality assurance tasks to test and evaluate the reasoning capabilities of LMMs. Our benchmark is curated from a variety of unusual and glitched scenarios from video games and aims to challenge both the visual and linguistic reasoning powers of LMMs in detecting and interpreting out-of-the-ordinary events. We evaluate multiple state-of-the-art LMMs and we show that GlitchBench presents a new challenge for these models. Code and data are available at: https://glitchbench.github.io/
****
Rethinking Interactive Image Segmentation with Low Latency High Quality and Diverse Prompts
Qin Liu, Jaemin Cho, Mohit Bansal, Marc Niethammer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3773-3782
The goal of interactive image segmentation is to delineate specific regions within an image via visual or language prompts. Low-latency and high-quality interactive segmentation with diverse prompts remain challenging for existing specialist and generalist models. Specialist models with their limited prompts and task-specific designs experience high latency because the image must be recomputed every time the prompt is updated due to the joint encoding of image and visual prompts. Generalist models exemplified by the Segment Anything Model (SAM) have recently excelled in prompt diversity and efficiency lifting image segmentation to the foundation model era. However for high-quality segmentations SAM still lags behind state-of-the-art specialist models despite SAM being trained with x100 more segmentation masks. In this work we delve deep into the architectural differences between the two types of models. We observe that dense representation and fusion of visual prompts are the key design choices contributing to the high segmentation quality of specialist models. In light of this we reintroduce this dense design into the generalist models to facilitate the development of generalist models with high segmentation quality. To densely represent diverse visual prompts we propose to use a dense map to capture five types: clicks boxes polygons scribbles and masks. Thus we propose SegNext a next-generation interactive segmentation approach offering low latency high quality and diverse prompt support. Our method outperforms current state-of-the-art methods on HQSeg-44K and DAVIS quantitatively and qualitatively.
****
ALGM: Adaptive Local-then-Global Token Merging for Efficient Semantic Segmentation with Plain Vision Transformers
Narges Norouzi, Svetlana Orlova, Daan de Geus, Gijs Dubbelman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15773-15782
This work presents Adaptive Local-then-Global Merging (ALGM) a token reduction method for semantic segmentation networks that use plain Vision Transformers. ALGM merges tokens in two stages: (1) In the first network layer it merges similar tokens within a small local window and (2) halfway through the network it merges similar tokens across the entire image. This is motivated by an analysis in which we found that in those situations tokens with a high cosine similarity can likely be merged without a drop in segmentation quality. With extensive experiments across multiple datasets and network configurations we show that ALGM not only significantly improves the throughput by up to 100% but can also enhance the mean IoU by up to +1.1 thereby achieving a better trade-off between segmentation quality and efficiency than existing methods. Moreover our approach is adaptive during inference meaning that the same model can be used for optimal efficiency or accuracy depending on the application. Code is available at https://tue-mps.github.io/ALGM.
****
DITTO: Dual and Integrated Latent Topologies for Implicit 3D Reconstruction
Jaehyeok Shim, Kyungdon Joo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5396-5405
We propose a novel concept of dual and integrated latent topologies (DITTO in short) for implicit 3D reconstruction from noisy and sparse point clouds. Most exi

sting methods predominantly focus on single latent type such as point or grid la
tents. In contrast the proposed DITTO leverages both point and grid latents (i.e
. dual latent) to enhance their strengths the stability of grid latents and the
detail-rich capability of point latents. Concretely DITTO consists of dual laten
t encoder and integrated implicit decoder. In the dual latent encoder a dual lat
ent layer which is the key module block composing the encoder refines both laten
ts in parallel maintaining their distinct shapes and enabling recursive interact
ion. Notably a newly proposed dynamic sparse point transformer within the dual l
atent layer effectively refines point latents. Then the integrated implicit deco
der systematically combines these refined latents achieving high-fidelity 3D rec
onstruction and surpassing previous state-of-the-art methods on object- and scen
e-level datasets especially in thin and detailed structures.
****
Single-Model and Any-Modality for Video Object Tracking
Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Dan
da Pani Paudel, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19156-19166
In the realm of video object tracking auxiliary modalities such as depth thermal
 or event data have emerged as valuable assets to complement the RGB trackers. I
n practice most existing RGB trackers learn a single set of parameters to use th
em across datasets and applications. However a similar single-model unification
for multi-modality tracking presents several challenges. These challenges stem f
rom the inherent heterogeneity of inputs -- each with modality-specific represen
tations the scarcity of multi-modal datasets and the absence of all the modaliti
es at all times. In this work we introduce Un-Track a Unified Tracker of a singl
e set of parameters for any modality. To handle any modality our method learns t
heir common latent space through low-rank factorization and reconstruction techn
iques. More importantly we use only the RGB-X pairs to learn the common latent s
pace. This unique shared representation seamlessly binds all modalities together
 enabling effective unification and accommodating any missing modality all withi
n a single transformer-based architecture. Our Un-Track achieves +8.1 absolute F
-score gain on the DepthTrack dataset by introducing only +2.14 (over 21.50) GFL
OPs with +6.6M (over 93M) parameters through a simple yet efficient prompting st
rategy. Extensive comparisons on five benchmark datasets with different modaliti
es show that Un-Track surpasses both SOTA unified trackers and modality-specific
 counterparts validating our effectiveness and practicality. The source code is
publicly available at https://github.com/Zongwei97/UnTrack.
****
FlowTrack: Revisiting Optical Flow for Long-Range Dense Tracking
Seokju Cho, Jiahui Huang, Seungryong Kim, Joon-Young Lee; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19
268-19277
In the domain of video tracking existing methods often grapple with a trade-off
between spatial density and temporal range. Current approaches in dense optical
flow estimators excel in providing spatially dense tracking but are limited to s
hort temporal spans. Conversely recent advancements in long-range trackers offer
 extended temporal coverage but at the cost of spatial sparsity. This paper intr
oduces FlowTrack a novel framework designed to bridge this gap. FlowTrack combin
es the strengths of both paradigms by 1) chaining confident flow predictions to
maximize efficiency and 2) automatically switching to an error compensation modu
le in instances of flow prediction inaccuracies. This dual strategy not only off
ers efficient dense tracking over extended temporal spans but also ensures robus
tness against error accumulations and occlusions common pitfalls of naive flow c
haining. Furthermore we demonstrate that chained flow itself can serve as an eff
ective guide for an error compensation module even for occluded points. Our fram
ework achieves state-of-the-art accuracy for long-range tracking on the DAVIS da
taset and renders 50% speed-up when performing dense tracking.
****
HIT: Estimating Internal Human Implicit Tissues from the Body Surface
Marilyn Keller, Vaibhav Arora, Abdelmouttaleb Dakri, Shivam Chandhok, Jürgen Mac

hann, Andreas Fritsche, Michael J. Black, Sergi Pujades; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3480-3490
The creation of personalized anatomical digital twins is important in the fields of medicine computer graphics sports science and biomechanics. To observe a subject's anatomy expensive medical devices (MRI or CT) are required and the creation of the digital model is often time-consuming and involves manual effort. Instead we leverage the fact that the shape of the body surface is correlated with the internal anatomy; e.g. from surface observations alone one can predict body composition and skeletal structure. In this work we go further and learn to infer the 3D location of three important anatomic tissues: subcutaneous adipose tissue (fat) lean tissue (muscles and organs) and long bones. To learn to infer these tissues we tackle several key challenges. We first create a dataset of human tissues by segmenting full-body MRI scans and registering the SMPL body mesh to the body surface. With this dataset we train HIT (Human Implicit Tissues) an implicit function that given a point inside a body predicts its tissue class. HIT leverages the SMPL body model shape and pose parameters to canonicalize the medical data. Unlike SMPL which is trained from upright 3D scans MRI scans are acquired with subjects lying on a table resulting in significant soft-tissue deformation. Consequently HIT uses a learned volumetric deformation field that undoes these deformations. Since HIT is parameterized by SMPL we can repose bodies or change the shape of subjects and the internal structures deform appropriately. We perform extensive experiments to validate HIT's ability to predict a plausible internal structure for novel subjects. The dataset and HIT model are available at https://hit.is.tue.mpg.de to foster future research in this direction.
****
DanceCamera3D: 3D Camera Movement Synthesis with Music and Dance
Zixuan Wang, Jia Jia, Shikun Sun, Haozhe Wu, Rong Han, Zhenyu Li, Di Tang, Jiaqing Zhou, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7892-7901
Choreographers determine what the dances look like while cameramen determine the final presentation of dances. Recently various methods and datasets have showcased the feasibility of dance synthesis. However camera movement synthesis with music and dance remains an unsolved challenging problem due to the scarcity of paired data. Thus we present DCM a new multi-modal 3D dataset which for the first time combines camera movement with dance motion and music audio. This dataset encompasses 108 dance sequences (3.2 hours) of paired dance-camera-music data from the anime community covering 4 music genres. With this dataset we uncover that dance camera movement is multifaceted and human-centric and possesses multiple influencing factors making dance camera synthesis a more challenging task compared to camera or dance synthesis alone. To overcome these difficulties we propose DanceCamera3D a transformer-based diffusion model that incorporates a novel body attention loss and a condition separation strategy. For evaluation we devise new metrics measuring camera movement quality diversity and dancer fidelity. Utilizing these metrics we conduct extensive experiments on our DCM dataset providing both quantitative and qualitative evidence showcasing the effectiveness of our DanceCamera3D model. Code and video demos are available at https://github.com/ Carmenw1203/DanceCamera3D-Official.
****
Synthesize Diagnose and Optimize: Towards Fine-Grained Vision-Language Understanding
Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, Zuxuan Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13279-13288
Vision language models (VLM) have demonstrated remarkable performance across various downstream tasks. However understanding fine-grained visual-linguistic concepts such as attributes and inter-object relationships remains a significant challenge. While several benchmarks aim to evaluate VLMs in finer granularity their primary focus remains on the linguistic aspect neglecting the visual dimension. Here we highlight the importance of evaluating VLMs from both a textual and vis

ual perspective. We introduce a progressive pipeline to synthesize images that vary in a specific attribute while ensuring consistency in all other aspects. Utilizing this data engine we carefully design a benchmark SPEC to diagnose the comprehension of object size position existence and count. Subsequently we conduct a thorough evaluation of four leading VLMs on SPEC. Surprisingly their performance is close to random guess revealing significant limitations. With this in mind we propose a simple yet effective approach to optimize VLMs in fine-grained understanding achieving significant improvements on SPEC without compromising the zero-shot performance. Results on two additional fine-grained benchmarks also show consistent improvements further validating the transferability of our approach. Code and data are available at https://github.com/wjpoom/SPEC.
****

Density-guided Translator Boosts Synthetic-to-Real Unsupervised Domain Adaptive Segmentation of 3D Point Clouds

Zhimin Yuan, Wankang Zeng, Yanfei Su, Weiquan Liu, Ming Cheng, Yulan Guo, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23303-23312

3D synthetic-to-real unsupervised domain adaptive segmentation is crucial to annotating new domains. Self-training is a competitive approach for this task but its performance is limited by different sensor sampling patterns (i.e. variations in point density) and incomplete training strategies. In this work we propose a density-guided translator (DGT) which translates point density between domains and integrates it into a two-stage self-training pipeline named DGT-ST. First in contrast to existing works that simultaneously conduct data generation and feature/output alignment within unstable adversarial training we employ the non-learnable DGT to bridge the domain gap at the input level. Second to provide a well-initialized model for self-training we propose a category-level adversarial network in stage one that utilizes the prototype to prevent negative transfer. Finally by leveraging the designs above a domain-mixed self-training method with source-aware consistency loss is proposed in stage two to narrow the domain gap further. Experiments on two synthetic-to-real segmentation tasks (SynLiDAR ? semanticKITTI and SynLiDAR ? semanticPOSS) demonstrate that DGT-ST outperforms state-of-the-art methods achieving 9.4% and 4.3% mIoU improvements respectively. Code is available at https://github.com/yuan-zm/DGT-ST.
****

Cross Initialization for Face Personalization of Text-to-Image Models

Lianyu Pang, Jian Yin, Haoran Xie, Qiping Wang, Qing Li, Xudong Mao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8393-8403

Recently there has been a surge in face personalization techniques benefiting from the advanced capabilities of pretrained text-to-image diffusion models. Among these a notable method is Textual Inversion which generates personalized images by inverting given images into textual embeddings. However methods based on Textual Inversion still struggle with balancing the trade-off between reconstruction quality and editability. In this study we examine this issue through the lens of initialization. Upon closely examining traditional initialization methods we identified a significant disparity between the initial and learned embeddings in terms of both scale and orientation. The scale of the learned embedding can be up to 100 times greater than that of the initial embedding. Such a significant change in the embedding could increase the risk of overfitting thereby compromising the editability. Driven by this observation we introduce a novel initialization method termed Cross Initialization that significantly narrows the gap between the initial and learned embeddings. This method not only improves both reconstruction and editability but also reduces the optimization steps from 5000 to 320. Furthermore we apply a regularization term to keep the learned embedding close to the initial embedding. We show that when combined with Cross Initialization this regularization term can effectively improve editability. We provide comprehensive empirical evidence to demonstrate the superior performance of our method compared to the baseline methods. Notably in our experiments Cross Initialization is the only method that successfully edits an individual's facial expression. A

dditionally a fast version of our method allows for capturing an input image in roughly 26 seconds while surpassing the baseline methods in terms of both reconstruction and editability. Code is available at https://github.com/lyuPang/CrossInitialization.
****
LEDITS++: Limitless Image Editing using Text-to-Image Models
Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, Apolinario Passos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8861-8870
Text-to-image diffusion models have recently received increasing interest for their astonishing ability to produce high-fidelity images from solely text inputs. Subsequent research efforts aim to exploit and apply their capabilities to real image editing. However existing image-to-image methods are often inefficient imprecise and of limited versatility. They either require time-consuming fine-tuning deviate unnecessarily strongly from the input image and/or lack support for multiple simultaneous edits. To address these issues we introduce LEDITS++ an efficient yet versatile and precise textual image manipulation technique. LEDITS++'s novel inversion approach requires no tuning nor optimization and produces high-fidelity results with a few diffusion steps. Second our methodology supports multiple simultaneous edits and is architecture-agnostic. Third we use a novel implicit masking technique that limits changes to relevant image regions. We propose the novel TEdBench++ benchmark as part of our exhaustive evaluation. Our results demonstrate the capabilities of LEDITS++ and its improvements over previous methods.
****
Video Interpolation with Diffusion Models
Siddhant Jain, Daniel Watson, Eric Tabellion, Aleksander Ho?ynski, Ben Poole, Janne Kontkanen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7341-7351
We present VIDIM a generative model for video interpolation which creates short videos given a start and end frame. In order to achieve high fidelity and generate motions unseen in the input data VIDIM uses cascaded diffusion models to first generate the target video at low resolution and then generate the high-resolution video conditioned on the low-resolution generated video. We compare VIDIM to previous state-of-the-art methods on video interpolation and demonstrate how such works fail in most settings where the underlying motion is complex nonlinear or ambiguous while VIDIM can easily handle such cases. We additionally demonstrate how classifier-free guidance on the start and end frame and conditioning the superresolution model on the original high-resolution frames without additional parameters unlocks high-fidelity results. VIDIM is fast to sample from as it jointly denoises all the frames to be generated requires less than a billion parameters per diffusion model to produce compelling results and still enjoys scalability and improved quality at larger parameter counts. Please see our project page at vidiminterpolation.github.io.
****
WildlifeMapper: Aerial Image Analysis for Multi-Species Detection and Identification
Satish Kumar, Bowen Zhang, Chandrakanth Gudavalli, Connor Levenson, Lacey Hughey, Jared A. Stabach, Irene Amoke, Gordon Ojwang, Joseph Mukeka, Stephen Mwiu, Joseph Ogutu, Howard Frederick, B.S. Manjunath; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12594-12604
We introduce WildlifeMapper (WM) a flexible model designed to detect locate and identify multiple species in aerial imagery. It addresses the limitations of traditional labor-intensive wildlife population assessments that are central to advancing environmental conservation efforts worldwide. While a number of methods exist to automate this process they are often limited in their ability to generalize to different species or landscapes due to the dominance of homogeneous backgrounds and/or poorly captured local image structures. WM introduces two novel modules that help to capture the local structure and context of objects of interest to accurately localize and identify them achieving a state-of-the-art (SOTA) d

etection rate of 0.56 mAP. Further we introduce a large aerial imagery dataset w
ith more than 11k Images and 28k annotations verified by trained experts. WM als
o achieves SOTA performance on 3 other publicly available aerial survey datasets
 collected across 4 different countries improving mAP by 42%. Source code and tr
ained models are available at Github
****
Learning Adaptive Spatial Coherent Correlations for Speech-Preserving Facial Exp
ression Manipulation
Tianshui Chen, Jianman Lin, Zhijing Yang, Chunmei Qing, Liang Lin; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 7267-7276
Speech-preserving facial expression manipulation (SPFEM) aims to modify facial e
motions while meticulously maintaining the mouth animation associated with spoke
n content. Current works depend on inaccessible paired training samples for the
person where two aligned frames exhibit the same speech content yet differ in em
otional expression limiting the SPFEM applications in real-world scenarios. In t
his work we discover that speakers who convey the same content with different em
otions exhibit highly correlated local facial animations providing valuable supe
rvision for SPFEM. To capitalize on this insight we propose a novel adaptive spa
tial coherent correlation learning (ASCCL) algorithm which models the aforementi
oned correlation as an explicit metric and integrates the metric to supervise ma
nipulating facial expression and meanwhile better preserving the facial animatio
n of spoken contents. To this end it first learns a spatial coherent correlation
 metric ensuring the visual disparities of adjacent local regions of the image b
elonging to one emotion are similar to those of the corresponding counterpart of
 the image belonging to another emotion. Recognizing that visual disparities are
 not uniform across all regions we have also crafted a disparity-aware adaptive
strategy that prioritizes regions that present greater challenges. During SPFEM
model training we construct the adaptive spatial coherent correlation metric bet
ween corresponding local regions of the input and output images as addition loss
 to supervise the generation process. We conduct extensive experiments on varian
t datasets and the results demonstrate the effectiveness of the proposed ASCCL a
lgorithm. Code is publicly available at https://github.com/jianmanlincjx/ASCCL
****
Tune-An-Ellipse: CLIP Has Potential to Find What You Want
Jinheng Xie, Songhe Deng, Bing Li, Haozhe Liu, Yawen Huang, Yefeng Zheng, Jurgen
 Schmidhuber, Bernard Ghanem, Linlin Shen, Mike Zheng Shou; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
13723-13732
Visual prompting of large vision language models such as CLIP exhibits intriguin
g zero-shot capabilities. A manually drawn red circle commonly used for highligh
ting can guide CLIP's attention to the surrounding region to identify specific o
bjects within an image. Without precise object proposals however it is insuffici
ent for localization. Our novel simple yet effective approach i.e. Differentiabl
e Visual Prompting enables CLIP to zero-shot localize: given an image and a text
 prompt describing an object we first pick a rendered ellipse from uniformly dis
tributed anchor ellipses on the image grid via visual prompting then use three l
oss functions to tune the ellipse coefficients to encapsulate the target region
gradually. This yields promising experimental results for referring expression c
omprehension without precisely specified object proposals. In addition we system
atically present the limitations of visual prompting inherent in CLIP and discus
s potential solutions.
****
Neural Spline Fields for Burst Image Fusion and Layer Separation
Ilya Chugunov, David Shustin, Ruyu Yan, Chenyang Lei, Felix Heide; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 25763-25773
Each photo in an image burst can be considered a sample of a complex 3D scene: t
he product of parallax diffuse and specular materials scene motion and illuminan
t variation. While decomposing all of these effects from a stack of misaligned i

mages is a highly ill-conditioned task the conventional align-and-merge burst pi
peline takes the other extreme: blending them into a single image. In this work
we propose a versatile intermediate representation: a two-layer alpha-composited
 image plus flow model constructed with neural spline fields -- networks trained
 to map input coordinates to spline control points. Our method is able to during
 test-time optimization jointly fuse a burst image capture into one high-resolut
ion reconstruction and decompose it into transmission and obstruction layers. Th
en by discarding the obstruction layer we can perform a range of tasks including
 seeing through occlusions reflection suppression and shadow removal. Tested on
complex in-the-wild captures we find that with no post-processing steps or learn
ed priors our generalizable model is able to outperform existing dedicated singl
e-image and multi-view obstruction removal approaches.
****

WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion
Soyong Shin, Juyong Kim, Eni Halilaj, Michael J. Black; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2070
-2080
The estimation of 3D human motion from video has progressed rapidly but current
methods still have several key limitations. First most methods estimate the huma
n in camera coordinates. Second prior work on estimating humans in global coordi
nates often assumes a flat ground plane and produces foot sliding. Third the mos
t accurate methods rely on computationally expensive optimization pipelines limi
ting their use to offline applications. Finally existing video-based methods are
 surprisingly less accurate than single-frame methods. We address these limitati
ons with WHAM (World-grounded Humans with Accurate Motion) which accurately and
efficiently reconstructs 3D human motion in a global coordinate system from vide
o. WHAM learns to lift 2D keypoint sequences to 3D using motion capture data and
 fuses this with video features integrating motion context and visual informatio
n. WHAM exploits camera angular velocity estimated from a SLAM method together w
ith human motion to estimate the body's global trajectory. We combine this with
a contact-aware trajectory refinement method that lets WHAM capture human motion
 in diverse conditions such as climbing stairs. WHAM outperforms all existing 3D
 human motion recovery methods across multiple in-the-wild benchmarks. Code is a
vailable for research purposes at http://wham.is.tue.mpg.de/.
****

NAPGuard: Towards Detecting Naturalistic Adversarial Patches
Siyang Wu, Jiakai Wang, Jiejie Zhao, Yazhe Wang, Xianglong Liu; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 24367-24376
Recently the emergence of naturalistic adversarial patch (NAP) which possesses a
 deceptive appearance and various representations underscores the necessity of d
eveloping robust detection strategies. However existing approaches fail to diffe
rentiate the deep-seated natures in adversarial patches i.e. aggressiveness and
naturalness leading to unsatisfactory precision and generalization against NAPs.
 To tackle this issue we propose NAPGuard to provide strong detection capability
 against NAPs via the elaborated critical feature modulation framework. For impr
oving precision we propose the aggressive feature aligned learning to enhance th
e model's capability in capturing accurate aggressive patterns. Considering the
challenge of inaccurate model learning caused by deceptive appearance we align t
he aggressive features by the proposed pattern alignment loss during training. S
ince the model could learn more accurate aggressive patterns it is able to detec
t deceptive patches more precisely. To enhance generalization we design the natu
ral feature suppressed inference to universally mitigate the disturbance from di
fferent NAPs. Since various representations arise in diverse disturbing forms to
 hinder generalization we suppress the natural features in a unified approach vi
a the feature shield module. Therefore the models could recognize NAPs within le
ss disturbance and activate the generalized detection ability. Extensive experim
ents show that our method surpasses state-of-the-art methods by large margins in
 detecting NAPs (improve 60.24% AP@0.5 on average).
****

DiffPerformer: Iterative Learning of Consistent Latent Guidance for Diffusion-based Human Video Generation
Chenyang Wang, Zerong Zheng, Tao Yu, Xiaoqian Lv, Bineng Zhong, Shengping Zhang, Liqiang Nie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6169-6179
Existing diffusion models for pose-guided human video generation mostly suffer from temporal inconsistency in the generated appearance and poses due to the inherent randomization nature of the generation process. In this paper we propose a novel framework DiffPerformer to synthesize high-fidelity and temporally consistent human video. Without complex architecture modification or costly training DiffPerformer finetunes a pretrained diffusion model on a single video of the target character and introduces an implicit video representation as a proxy to learn temporally consistent guidance for the diffusion model. The guidance is encoded into VAE latent space and an iterative optimization loop is constructed between the implicit video representation and the diffusion model allowing to harness the smooth property of the implicit video representation and the generative capabilities of the diffusion model in a mutually beneficial way. Moreover we propose 3D-aware human flow as a temporal constraint during the optimization to explicitly model the correspondence between driving poses and human appearance. This alleviates the misalignment between guided poses and target performer and therefore maintains the appearance coherence under various motions. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods.
****
Unified Language-driven Zero-shot Domain Adaptation
Senqiao Yang, Zhuotao Tian, Li Jiang, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23407-23415
This paper introduces Unified Language-driven Zero-shot Domain Adaptation (ULDA) a novel task setting that enables a single model to adapt to diverse target domains without explicit domain-ID knowledge. We identify the constraints in the existing language-driven zero-shot domain adaptation task particularly the requirement for domain IDs and domain-specific models which may restrict flexibility and scalability. To overcome these issues we propose a new framework for ULDA consisting of Hierarchical Context Alignment (HCA) Domain Consistent Representation Learning (DCRL) and Text-Driven Rectifier (TDR). These components work synergistically to align simulated features with target text across multiple visual levels retain semantic correlations between different regional representations and rectify biases between simulated and real target visual features respectively. Our extensive empirical evaluations demonstrate that this framework achieves competitive performance in both settings surpassing even the model that requires domain-ID showcasing its superiority and generalization ability. The proposed method is not only effective but also maintains practicality and efficiency as it does not introduce additional computational costs during inference. The code is available on the project website.
****
Category-Level Multi-Part Multi-Joint 3D Shape Assembly
Yichen Li, Kaichun Mo, Yueqi Duan, He Wang, Jiequan Zhang, Lin Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3281-3291
Shape assembly composes complex shapes geometries by arranging simple part geometries and has wide applications in autonomous robotic assembly and CAD modeling. Existing works focus on geometry reasoning and neglect the actual physical assembly process of matching and fitting joints which are the contact surfaces connecting different parts. In this paper we consider contacting joints for the task of multi-part assembly. A successful joint-optimized assembly needs to satisfy the bilateral objectives of shape structure and joint alignment. We propose a hierarchical graph learning approach composed of two levels of graph representation learning. The part graph takes part geometries as input to build the desired shape structure. The joint-level graph uses part joints information and focuses on matching and aligning joints. The two kinds of information are combined to achieve the bilateral objectives. Extensive experiments demonstrate that our method

outperforms previous methods achieving better shape structure and higher joint alignment accuracy.
****
Equivariant Multi-Modality Image Fusion
Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25912-25921
Multi-modality image fusion is a technique that combines information from different sensors or modalities enabling the fused image to retain complementary features from each modality such as functional highlights and texture details. However effective training of such fusion models is challenging due to the scarcity of ground truth fusion data. To tackle this issue we propose the Equivariant Multi-Modality imAge fusion (EMMA) paradigm for end-to-end self-supervised learning. Our approach is rooted in the prior knowledge that natural imaging responses are equivariant to certain transformations. Consequently we introduce a novel training paradigm that encompasses a fusion module a pseudo-sensing module and an equivariant fusion module. These components enable the net training to follow the principles of the natural sensing-imaging process while satisfying the equivariant imaging prior. Extensive experiments confirm that EMMA yields high-quality fusion results for infrared-visible and medical images concurrently facilitating downstream multi-modal segmentation and detection tasks. The code is available at https://github.com/Zhaozixiang1228/MMIF-EMMA.
****
NeLF-Pro: Neural Light Field Probes for Multi-Scale Novel View Synthesis
Zinuo You, Andreas Geiger, Anpei Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19833-19843
We present NeLF-Pro a novel representation to model and reconstruct light fields in diverse natural scenes that vary in extent and spatial granularity. In contrast to previous fast reconstruction methods that represent the 3D scene globally we model the light field of a scene as a set of local light field feature probes parameterized with position and multi-channel 2D feature maps. Our central idea is to bake the scene's light field into spatially varying learnable representations and to query point features by weighted blending of probes close to the camera - allowing for mipmap representation and rendering. We introduce a novel vector-matrix-matrix (VMM) factorization technique that effectively represents the light field feature probes as products of core factors (i.e. VM) shared among local feature probes and a basis factor (i.e. M) - efficiently encoding internal relationships and patterns within the scene.Experimentally we demonstrate that NeLF-Pro significantly boosts the performance of feature grid-based representations and achieves fast reconstruction with better rendering quality while maintaining compact modeling. Project page: sinoyou.github.io/nelf-pro
****
One-Shot Open Affordance Learning with Foundation Models
Gen Li, Deqing Sun, Laura Sevilla-Lara, Varun Jampani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3086-3096
We introduce One-shot Open Affordance Learning (OOAL) where a model is trained with just one example per base object category but is expected to identify novel objects and affordances. While vision-language models excel at recognizing novel objects and scenes they often struggle to understand finer levels of granularity such as affordances. To handle this issue we conduct a comprehensive analysis of existing foundation models to explore their inherent understanding of affordances and assess the potential for data-limited affordance learning. We then propose a vision-language framework with simple and effective designs that boost the alignment between visual features and affordance text embeddings. Experiments on two affordance segmentation benchmarks show that the proposed method outperforms state-of-the-art models with less than 1% of the full training data and exhibits reasonable generalization capability on unseen objects and affordances. Project page: https://reagan1311.github.io/ooal.
****

Don't Look into the Dark: Latent Codes for Pluralistic Image Inpainting
Haiwei Chen, Yajie Zhao; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 7591-7600
We present a method for large-mask pluralistic image inpainting based on the gen
erative framework of discrete latent codes. Our method learns latent priors disc
retized as tokens by only performing computations at the visible locations of th
e image. This is realized by a restrictive partial encoder that predicts the tok
en label for each visible block a bidirectional transformer that infers the miss
ing labels by only looking at these tokens and a dedicated synthesis network tha
t couples the tokens with the partial image priors to generate coherent and plur
alistic complete image even under extreme mask settings. Experiments on public b
enchmarks validate our design choices as the proposed method outperforms strong
baselines in both visual quality and diversity metrics.
****
Incremental Nuclei Segmentation from Histopathological Images via Future-class A
wareness and Compatibility-inspired Distillation
Huyong Wang, Huisi Wu, Jing Qin; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2024, pp. 11408-11417
We present a novel semantic segmentation approach for incremental nuclei segment
ation from histopathological images which is a very challenging task as we have
to incrementally optimize existing models to make them perform well in both old
and new classes without using training samples of old classes. Yet it is an indi
spensable component of computer-aided diagnosis systems. The proposed approach h
as two key techniques. First we propose a new future-class awareness mechanism b
y separating some potential regions for future classes from background based on
their similarities to both old and new classes in the representation space. With
 this mechanism we can not only reserve more parameter space for future updates
but also enhance the representation capability of learned features. We further p
ropose an innovative compatibility-inspired distillation scheme to make our mode
l take full advantage of the knowledge learned by the old model. We conducted ex
tensive experiments on two famous histopathological datasets and the results dem
onstrate the proposed approach achieves much better performance than state-of-th
e-art approaches. The code is available at https://github.com/why19991/InSeg.
****
DiffEditor: Boosting Accuracy and Flexibility on Diffusion-based Image Editing
Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, Jian Zhang; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 8488-8497
Large-scale Text-to-Image (T2I) diffusion models have revolutionized image gener
ation over the last few years. Although owning diverse and high-quality generati
on capabilities translating these abilities to fine-grained image editing remain
s challenging. In this paper we propose DiffEditor to rectify two weaknesses in
existing diffusion-based image editing: (1) in complex scenarios editing results
 often lack editing accuracy and exhibit unexpected artifacts; (2) lack of flexi
bility to harmonize editing operations e.g. imagine new content. In our solution
 we introduce image prompts in fine-grained image editing cooperating with the t
ext prompt to better describe the editing content. To increase the flexibility w
hile maintaining content consistency we locally combine stochastic differential
equation (SDE) into the ordinary differential equation (ODE) sampling. In additi
on we incorporate regional score-based gradient guidance and a time travel strat
egy into the diffusion sampling further improving the editing quality. Extensive
 experiments demonstrate that our method can efficiently achieve state-of-the-ar
t performance on various fine-grained image editing tasks including editing with
in a single image (e.g. object moving resizing and content dragging) and across
images (e.g. appearance replacing and object pasting). Our source code is releas
ed at https://github.com/MC-E/DragonDiffusion.
****
Solving Masked Jigsaw Puzzles with Diffusion Vision Transformers
Jinyang Liu, Wondmgezahu Teshome, Sandesh Ghimire, Mario Sznaier, Octavia Camps;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti

on (CVPR), 2024, pp. 23009-23018
Solving image and video jigsaw puzzles poses the challenging task of rearranging image fragments or video frames from unordered sequences to restore meaningful images and video sequences. Existing approaches often hinge on discriminative mo dels tasked with predicting either the absolute positions of puzzle elements or the permutation actions applied to the original data. Unfortunately these method s face limitations in effectively solving puzzles with a large number of element s. In this paper we propose JPDVT an innovative approach that harnesses diffusio n transformers to address this challenge. Specifically we generate positional in formation for image patches or video frames conditioned on their underlying visu al content. This information is then employed to accurately assemble the puzzle pieces in their correct positions even in scenarios involving missing pieces. Ou r method achieves state-of-the-art performance on several datasets.
****
InstructVideo: Instructing Video Diffusion Models with Human Feedback
Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, Dong Ni; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6463-6474
Diffusion models have emerged as the de facto paradigm for video generation. How ever their reliance on web-scale data of varied quality often yields results tha t are visually unappealing and misaligned with the textual prompts. To tackle th is problem we propose InstructVideo to instruct text-to-video diffusion models w ith human feedback by reward fine-tuning. InstructVideo has two key ingredients: 1) To ameliorate the cost of reward fine-tuning induced by generating through t he full DDIM sampling chain we recast reward fine-tuning as editing. By leveragi ng the diffusion process to corrupt a sampled video InstructVideo requires only partial inference of the DDIM sampling chain reducing fine-tuning cost while imp roving fine-tuning efficiency. 2) To mitigate the absence of a dedicated video r eward model for human preferences we repurpose established image reward models e .g. HPSv2. To this end we propose Segmental Video Reward a mechanism to provide reward signals based on segmental sparse sampling and Temporally Attenuated Rewa rd a method that mitigates temporal modeling degradation during fine-tuning. Ext ensive experiments both qualitative and quantitative validate the practicality a nd efficacy of using image reward models in InstructVideo significantly enhancin g the visual quality of generated videos without compromising generalization cap abilities. Code and models can be accessed through our project page https://inst ructvideo.github.io/.
****
Fully Exploiting Every Real Sample: SuperPixel Sample Gradient Model Stealing
Yunlong Zhao, Xiaoheng Deng, Yijing Liu, Xinjun Pei, Jiazhi Xia, Wei Chen; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2024, pp. 24316-24325
Model stealing (MS) involves querying and observing the output of a machine lear ning model to steal its capabilities. The quality of queried data is crucial yet obtaining a large amount of real data for MS is often challenging. Recent works have reduced reliance on real data by using generative models. However when hig h-dimensional query data is required these methods are impractical due to the hi gh costs of querying and the risk of model collapse. In this work we propose usi ng sample gradients (SG) to enhance the utility of each real sample as SG provid es crucial guidance on the decision boundaries of the victim model. However util izing SG in the model stealing scenario faces two challenges: 1. Pixel-level gra dient estimation requires extensive query volume and is susceptible to defenses. 2. The estimation of sample gradients has a significant variance. This paper pr oposes Superpixel Sample Gradient stealing (SPSG) for model stealing under the c onstraint of limited real samples. With the basic idea of imitating the victim m odel's low-variance patch-level gradients instead of pixel-level gradients SPSG achieves efficient sample gradient estimation through two steps. First we perfor m patch-wise perturbations on query images to estimate the average gradient in d ifferent regions of the image. Then we filter the gradients through a threshold strategy to reduce variance. Exhaustive experiments demonstrate that with the sa

me number of real samples SPSG achieves accuracy agreements and adversarial succ
ess rate significantly surpassing the current state-of-the-art MS methods. Codes
 are available at https://github.com/zyl123456aB/SPSG_attack.
****
Progressive Divide-and-Conquer via Subsampling Decomposition for Accelerated MRI
Chong Wang, Lanqing Guo, Yufei Wang, Hao Cheng, Yi Yu, Bihan Wen; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 25128-25137
Deep unfolding networks (DUN) have emerged as a popular iterative framework for
accelerated magnetic resonance imaging (MRI) reconstruction. However conventiona
l DUN aims to reconstruct all the missing information within the entire space in
 each iteration. Thus it could be challenging when dealing with highly ill-posed
 degradation often resulting in subpar reconstruction. In this work we propose a
 Progressive Divide-And-Conquer (PDAC) strategy aiming to break down the subsamp
ling process in the actual severe degradation and thus perform reconstruction se
quentially. Starting from decomposing the original maximum-a-posteriori problem
of accelerated MRI we present a rigorous derivation of the proposed PDAC framewo
rk which could be further unfolded into an end-to-end trainable network. Each PD
AC iteration specifically targets a distinct segment of moderate degradation bas
ed on the decomposition. Furthermore as part of the PDAC iteration such decompos
ition is adaptively learned as an auxiliary task through a degradation predictor
 which provides an estimation of the decomposed sampling mask. Following this pr
ediction the sampling mask is further integrated via a severity conditioning mod
ule to ensure awareness of the degradation severity at each stage. Extensive exp
eriments demonstrate that our proposed method achieves superior performance on t
he publicly available fastMRI and Stanford2D FSE datasets in both multi-coil and
 single-coil settings.
****
DiffMOT: A Real-time Diffusion-based Multiple Object Tracker with Non-linear Pre
diction
Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, Dan Zeng; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 19321-19330
In Multiple Object Tracking objects often exhibit non-linear motion of accelerat
ion and deceleration with irregular direction changes. Tacking-by-detection (TBD
) trackers with Kalman Filter motion prediction work well in pedestrian-dominant
 scenarios but fall short in complex situations when multiple objects perform no
n-linear and diverse motion simultaneously. To tackle the complex non-linear mot
ion we propose a real-time diffusion-based MOT approach named DiffMOT. Specifica
lly for the motion predictor component we propose a novel Decoupled Diffusion-ba
sed Motion Predictor (D^2MP). It models the entire distribution of various motio
n presented by the data as a whole. It also predicts an individual object's moti
on conditioning on an individual's historical motion information. Furthermore it
 optimizes the diffusion process with much fewer sampling steps. As a MOT tracke
r the DiffMOT is real-time at 22.7FPS and also outperforms the state-of-the-art
on DanceTrack and SportsMOT datasets with 62.3% and 76.2% in HOTA metrics respec
tively. To the best of our knowledge DiffMOT is the first to introduce a diffusi
on probabilistic model into the MOT to tackle non-linear motion prediction.
****
MV-Adapter: Multimodal Video Transfer Learning for Video Text Retrieval
Xiaojie Jin, Bowen Zhang, Weibo Gong, Kai Xu, Xueqing Deng, Peng Wang, Zhao Zhan
g, Xiaohui Shen, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2024, pp. 27144-27153
State-of-the-art video-text retrieval (VTR) methods typically involve fully fine
-tuning a pre-trained model (e.g. CLIP) on specific datasets. However this can r
esult in significant storage costs in practical applications as a separate model
 per task must be stored. To address this issue we present our pioneering work t
hat enables parameter-efficient VTR using a pre-trained model with only a small
number of tunable parameters during training. Towards this goal we propose a new
 method dubbed Multimodal Video Adapter (MV-Adapter) for efficiently transferrin

g the knowledge in the pre-trained CLIP from image-text to video-text. Specifically MV-Adapter utilizes bottleneck structures in both video and text branches along with two novel components. The first is a Temporal Adaptation Module that is incorporated in the video branch to introduce global and local temporal contexts. We also train weights calibrations to adjust to dynamic variations across frames. The second is Cross Modality Tying that generates weights for video/text branches through sharing cross modality factors for better aligning between modalities. Thanks to above innovations MV-Adapter can achieve comparable or better performance than standard fine-tuning with negligible parameters overhead. Notably MV-Adapter consistently outperforms various competing methods in V2T/T2V tasks with large margins on five widely used VTR benchmarks (MSR-VTT MSVD LSMDC DiDemo and ActivityNet). Codes will be released.
****
Rethinking Multi-view Representation Learning via Distilled Disentangling
Guanzhou Ke, Bo Wang, Xiaoli Wang, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26774-26783
Multi-view representation learning aims to derive robust representations that are both view-consistent and view-specific from diverse data sources. This paper presents an in-depth analysis of existing approaches in this domain highlighting a commonly overlooked aspect: the redundancy between view-consistent and view-specific representations. To this end we propose an innovative framework for multi-view representation learning which incorporates a technique we term 'distilled disentangling'. Our method introduces the concept of masked cross-view prediction enabling the extraction of compact high-quality view-consistent representations from various sources without incurring extra computational overhead. Additionally we develop a distilled disentangling module that efficiently filters out consistency-related information from multi-view representations resulting in purer view-specific representations. This approach significantly reduces redundancy between view-consistent and view-specific representations enhancing the overall efficiency of the learning process. Our empirical evaluations reveal that higher mask ratios substantially improve the quality of view-consistent representations. Moreover we find that reducing the dimensionality of view-consistent representations relative to that of view-specific representations further refines the quality of the combined representations.
****
Just Add ?! Pose Induced Video Transformers for Understanding Activities of Daily Living
Dominick Reilly, Srijan Das; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18340-18350
Video transformers have become the de facto standard for human action recognition yet their exclusive reliance on the RGB modality still limits their adoption in certain domains. One such domain is Activities of Daily Living (ADL) where RGB alone is not sufficient to distinguish between visually similar actions or actions observed from multiple viewpoints. To facilitate the adoption of video transformers for ADL we hypothesize that the augmentation of RGB with human pose information known for its sensitivity to fine-grained motion and multiple viewpoints is essential. Consequently we introduce the first Pose Induced Video Transformer: PI-ViT (or ?-ViT) a novel approach that augments the RGB representations learned by video transformers with 2D and 3D pose information. The key elements of ?-ViT are two plug-in modules 2D Skeleton Induction Module and 3D Skeleton Induction Module that are responsible for inducing 2D and 3D pose information into the RGB representations. These modules operate by performing pose-aware auxiliary tasks a design choice that allows ?-ViT to discard the modules during inference. Notably ?-ViT achieves the state-of-the-art performance on three prominent ADL datasets encompassing both real-world and large-scale RGB-D datasets without requiring poses or additional computational overhead at inference.
****
ViLa-MIL: Dual-scale Vision-Language Multiple Instance Learning for Whole Slide Image Classification
Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, Huazhu Fu; Proceedings of the

 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 11248-11258
Multiple instance learning (MIL)-based framework has become the mainstream for p
rocessing the whole slide image (WSI) with giga-pixel size and hierarchical imag
e context in digital pathology. However these methods heavily depend on a substa
ntial number of bag-level labels and solely learn from the original slides which
 are easily affected by variations in data distribution. Recently vision languag
e model (VLM)-based methods introduced the language prior by pre-training on lar
ge-scale pathological image-text pairs. However the previous text prompt lacks t
he consideration of pathological prior knowledge therefore does not substantiall
y boost the model's performance. Moreover the collection of such pairs and the p
re-training process are very time-consuming and source-intensive. To solve the a
bove problems we propose a dual-scale vision-language multiple instance learning
 (ViLa-MIL) framework for whole slide image classification. Specifically we prop
ose a dual-scale visual descriptive text prompt based on the frozen large langua
ge model (LLM) to boost the performance of VLM effectively. To transfer the VLM
to process WSI efficiently for the image branch we propose a prototype-guided pa
tch decoder to aggregate the patch features progressively by grouping similar pa
tches into the same prototype; for the text branch we introduce a context-guided
 text decoder to enhance the text features by incorporating the multi-granular i
mage contexts. Extensive studies on three multi-cancer and multi-center subtypin
g datasets demonstrate the superiority of ViLa-MIL.
****
Targeted Representation Alignment for Open-World Semi-Supervised Learning
Ruixuan Xiao, Lei Feng, Kai Tang, Junbo Zhao, Yixuan Li, Gang Chen, Haobo Wang;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2024, pp. 23072-23082
Open-world Semi-Supervised Learning aims to classify unlabeled samples utilizing
 information from labeled data while unlabeled samples are not only from the lab
eled known categories but also from novel categories previously unseen. Despite
the promise current approaches solely rely on hazardous similarity-based cluster
ing algorithms and give unlabeled samples free rein to spontaneously group into
distinct novel class clusters. Nevertheless due to the absence of novel class su
pervision these methods typically suffer from the representation collapse dilemm
a---features of different novel categories can get closely intertwined and indis
tinguishable even collapsing into the same cluster and leading to degraded perfo
rmance. To alleviate this we propose a novel framework TRAILER which targets to
attain an optimal feature arrangement revealed by the recently uncovered neural
collapse phenomenon. To fulfill this we adopt targeted prototypes that are pre-a
ssigned uniformly with maximum separation and then progressively align the repre
sentations to them. To further tackle the potential downsides of such stringent
alignment we encapsulate a sample-target allocation mechanism with coarse-to-fin
e refinery that is able to infer label assignments with high quality. Extensive
experiments demonstrate that TRAILER outperforms current state-of-the-art method
s on generic and fine-grained benchmarks. The code is available at https://githu
b.com/Justherozen/TRAILER.
****
Efficient Solution of Point-Line Absolute Pose
Petr Hruby, Timothy Duff, Marc Pollefeys; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21316-21325
We revisit certain problems of pose estimation based on 3D--2D correspondences b
etween features which may be points or lines. Specifically we address the two pr
eviously-studied minimal problems of estimating camera extrinsics from p \in \
1 2 \ point--point correspondences and l=3-p line--line correspondences. To the
 best of our knowledge all of the previously-known practical solutions to these
problems required computing the roots of degree \ge 4 (univariate) polynomials w
hen p=2 or degree \ge 8 polynomials when p=1. We describe and implement two elem
entary solutions which reduce the degrees of the needed polynomials from 4 to 2
and from 8 to 4 respectively. We show experimentally that the resulting solvers
are numerically stable and fast: when compared to the previous state-of-the art

we may obtain nearly an order of magnitude speedup. The code is available at htt
ps://github.com/petrhruby97/efficient_absolute
****
Text-to-3D using Gaussian Splatting
Zilong Chen, Feng Wang, Yikai Wang, Huaping Liu; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21401-21412
Automatic text-to-3D generation that combines Score Distillation Sampling (SDS)
with the optimization of volume rendering has achieved remarkable progress in sy
nthesizing realistic 3D objects. Yet most existing text-to-3D methods by SDS and
 volume rendering suffer from inaccurate geometry e.g. the Janus issue since it
is hard to explicitly integrate 3D priors into implicit 3D representations. Besi
des it is usually time-consuming for them to generate elaborate 3D models with r
ich colors. In response this paper proposes GSGEN a novel method that adopts Gau
ssian Splatting a recent state-of-the-art representation to text-to-3D generatio
n. GSGEN aims at generating high-quality 3D objects and addressing existing shor
tcomings by exploiting the explicit nature of Gaussian Splatting that enables th
e incorporation of 3D prior. Specifically our method adopts a progressive optimi
zation strategy which includes a geometry optimization stage and an appearance r
efinement stage. In geometry optimization a coarse representation is established
 under 3D point cloud diffusion prior along with the ordinary 2D SDS optimizatio
n ensuring a sensible and 3D-consistent rough shape. Subsequently the obtained G
aussians undergo an iterative appearance refinement to enrich texture details. I
n this stage we increase the number of Gaussians by compactness-based densificat
ion to enhance continuity and improve fidelity. With these designs our approach
can generate 3D assets with delicate details and accurate geometry. Extensive ev
aluations demonstrate the effectiveness of our method especially for capturing h
igh-frequency components.
****
CapsFusion: Rethinking Image-Text Data at Scale
Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wan
g, Jingjing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2024, pp. 14022-14032
Large multimodal models demonstrate remarkable generalist ability to perform div
erse multimodal tasks in a zero-shot manner. Large-scale web-based image-text pa
irs contribute fundamentally to this success but suffer from excessive noise. Re
cent studies use alternative captions synthesized by captioning models and have
achieved notable benchmark performance. However our experiments reveal significa
nt Scalability Deficiency and World Knowledge Loss issues in models trained with
 synthetic captions which have been largely obscured by their initial benchmark
success. Upon closer examination we identify the root cause as the overly-simpli
fied language structure and lack of knowledge details in existing synthetic capt
ions. To provide higher-quality and more scalable multimodal pretraining data we
 propose CapsFusion an advanced framework that leverages large language models t
o consolidate and refine information from both web-based image-text pairs and sy
nthetic captions. Extensive experiments show that CapsFusion captions exhibit re
markable all-round superiority over existing captions in terms of model performa
nce (e.g. 18.8 and 18.3 improvements in CIDEr score on COCO and NoCaps) sample e
fficiency (requiring 11-16 times less computation than baselines) world knowledg
e depth and scalability. These effectiveness efficiency and scalability advantag
es position CapsFusion as a promising candidate for future scaling of LMM traini
ng.
****
On the Content Bias in Frechet Video Distance
Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, Jia-Bin Huang; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 7277-7288
Frechet Video Distance (FVD) a prominent metric for evaluating video generation
models is known to conflict with human perception occasionally. In this paper we
 aim to explore the extent of FVD's bias toward frame quality over temporal real
ism and identify its sources. We first quantify the FVD's sensitivity to the tem

poral axis by decoupling the frame and motion quality and find that the FVD only increases slightly with larger temporal corruption. We then analyze the generated videos and show that via careful sampling from a large set of generated videos that do not contain motions one can drastically decrease FVD without improving the temporal quality. Both studies suggest FVD's basis towards the quality of individual frames. We show that FVD with features extracted from the recent large-scale self-supervised video models is less biased toward image quality. Finally we revisit a few real-world examples to validate our hypothesis.
****
Tumor Micro-environment Interactions Guided Graph Learning for Survival Analysis of Human Cancers from Whole-slide Pathological Images
Wei Shao, YangYang Shi, Daoqiang Zhang, JunJie Zhou, Peng Wan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11694-11703
The recent advance of deep learning technology brings the possibility of assisting the pathologist to predict the patients' survival from whole-slide pathological images (WSIs). However most of the prevalent methods only worked on the sampled patches in specifically or randomly selected tumor areas of WSIs which has very limited capability to capture the complex interactions between tumor and its surrounding micro-environment components. As a matter of fact tumor is supported and nurtured in the heterogeneous tumor micro-environment(TME) and the detailed analysis of TME and their correlation with tumors are important to in-depth analyze the mechanism of cancer development. In this paper we considered the spatial interactions among tumor and its two major TME components (i.e. lymphocytes and stromal fibrosis) and presented a Tumor Micro-environment Interactions Guided Graph Learning (TMEGL) algorithm for the prognosis prediction of human cancers. Specifically we firstly selected different types of patches as nodes to build graph for each WSI. Then a novel TME neighborhood organization guided graph embedding algorithm was proposed to learn node representations that can preserve their topological structure information. Finally a Gated Graph Attention Network is applied to capture the survival-associated intersections among tumor and different TME components for clinical outcome prediction. We tested TMEGL on three cancer cohorts derived from The Cancer Genome Atlas (TCGA) and the experimental results indicated that TMEGL not only outperforms the existing WSI-based survival analysis models but also has good explainable ability for survival prediction.
****
Towards Generalizable Multi-Object Tracking
Zheng Qin, Le Wang, Sanping Zhou, Panpan Fu, Gang Hua, Wei Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18995-19004
Multi-Object Tracking (MOT) encompasses various tracking scenarios each characterized by unique traits. Effective trackers should demonstrate a high degree of generalizability across diverse scenarios. However existing trackers struggle to accommodate all aspects or necessitate hypothesis and experimentation to customize the association information (motion and/or appearance) for a given scenario leading to narrowly tailored solutions with limited generalizability. In this paper we investigate the factors that influence trackers' generalization to different scenarios and concretize them into a set of tracking scenario attributes to guide the design of more generalizable trackers. Furthermore we propose a "point-wise to instance-wise relation" framework for MOT i.e. GeneralTrack which can generalize across diverse scenarios while eliminating the need to balance motion and appearance. Thanks to its superior generalizability our proposed GeneralTrack achieves state-of-the-art performance on multiple benchmarks and demonstrates the potential for domain generalization.
****
POPDG: Popular 3D Dance Generation with PopDanceSet
Zhenye Luo, Min Ren, Xuecai Hu, Yongzhen Huang, Li Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26984-26993
Generating dances that are both lifelike and well-aligned with music continues t

o be a challenging task in the cross-modal domain. This paper introduces PopDanc eSet the first dataset tailored to the preferences of young audiences enabling t he generation of aesthetically oriented dances. And it surpasses the AIST++ data set in music genre diversity and the intricacy and depth of dance movements. Mor eover the proposed POPDG model within the iDDPM framework enhances dance diversi ty and through the Space Augmentation Algorithm strengthens spatial physical con nections between human body joints ensuring that increased diversity does not co mpromise generation quality. A streamlined Alignment Module is also designed to improve the temporal alignment between dance and music. Extensive experiments sh ow that POPDG achieves SOTA results on two datasets. Furthermore the paper also expands on current evaluation metrics. The dataset and code are available at htt ps://github.com/Luke-Luo1/POPDG.
****
Image Neural Field Diffusion Models
Yinbo Chen, Oliver Wang, Richard Zhang, Eli Shechtman, Xiaolong Wang, Michael Gh arbi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2024, pp. 8007-8017
Diffusion models have shown an impressive ability to model complex data distribu tions with several key advantages over GANs such as stable training better cover age of the training distribution's modes and the ability to solve inverse proble ms without extra training. However most diffusion models learn the distribution of fixed-resolution images. We propose to learn the distribution of continuous i mages by training diffusion models on image neural fields which can be rendered at any resolution and show its advantages over fixed-resolution models. To achie ve this a key challenge is to obtain a latent space that represents photorealist ic image neural fields. We propose a simple and effective method inspired by sev eral recent techniques but with key changes to make the image neural fields phot orealistic. Our method can be used to convert existing latent diffusion autoenco ders into image neural field autoencoders. We show that image neural field diffu sion models can be trained using mixed-resolution image datasets outperform fixe d-resolution diffusion models followed by super-resolution models and can solve inverse problems with conditions applied at different scales efficiently.
****
Discriminative Probing and Tuning for Text-to-Image Generation
Leigang Qu, Wenjie Wang, Yongqi Li, Hanwang Zhang, Liqiang Nie, Tat-Seng Chua; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7434-7444
Despite advancements in text-to-image generation (T2I) prior methods often face text-image misalignment problems such as relation confusion in generated images. Existing solutions involve cross-attention manipulation for better compositiona l understanding or integrating large language models for improved layout plannin g. However the inherent alignment capabilities of T2I models are still inadequat e. By reviewing the link between generative and discriminative modeling we posit that T2I models' discriminative abilities may reflect their text-image alignmen t proficiency during generation. In this light we advocate bolstering the discri minative abilities of T2I models to achieve more precise text-to-image alignment for generation. We present a discriminative adapter built on T2I models to prob e their discriminative abilities on two representative tasks and leverage discri minative fine-tuning to improve their text-image alignment. As a bonus of the di scriminative adapter a self-correction mechanism can leverage discriminative gra dients to better align generated images to text prompts during inference. Compre hensive evaluations across three benchmark datasets including both in-distributi on and out-of-distribution scenarios demonstrate our method's superior generatio n performance. Meanwhile it achieves state-of-the-art discriminative performance on the two discriminative tasks compared to other generative models. The code i s available at https://dpt-t2i.github.io/.
****
Slice3D: Multi-Slice Occlusion-Revealing Single View 3D Reconstruction
Yizhi Wang, Wallace Lira, Wenqi Wang, Ali Mahdavi-Amiri, Hao Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20

24, pp. 9881-9891
We introduce multi-slice reasoning a new notion for single-view 3D reconstructio n which challenges the current and prevailing belief that multi-view synthesis i s the most natural conduit between single-view and 3D. Our key observation is th at object slicing is a more direct and hence more advantageous means to reveal o ccluded structures than altering camera views. Specifically slicing can peel thr ough any occluder without obstruction and in the limit (i.e. with infinitely man y slices) it is guaranteed to unveil all hidden object parts. We realize our ide a by developing Slice3D a novel method for single-view 3D reconstruction which f irst predicts multi-slice images from a single RGB input image and then integrat es the slices into a 3D model using a coordinate-based transformer network to pr oduct a signed distance function. The slice images can be regressed or generated both through a U-Net based network. For the former we inject a learnable slice indicator code to designate each decoded image into a spatial slice location whi le the slice generator is a denoising diffusion model operating on the entirety of slice images stacked on the input channels. We conduct extensive evaluation a gainst state-of-the-art alternatives to demonstrate superiority of our method es pecially in recovering complex and severely occluded shape structures amid ambig uities. All Slice3D results were produced by networks trained on a single Nvidia A40 GPU with an inference time of less than 20 seconds.
****
Towards More Accurate Diffusion Model Acceleration with A Timestep Tuner
Mengfei Xia, Yujun Shen, Changsong Lei, Yu Zhou, Deli Zhao, Ran Yi, Wenping Wang , Yong-Jin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2024, pp. 5736-5745
A diffusion model which is formulated to produce an image using thousands of den oising steps usually suffers from a slow inference speed. Existing acceleration algorithms simplify the sampling by skipping most steps yet exhibit considerable performance degradation. By viewing the generation of diffusion models as a dis cretized integral process we argue that the quality drop is partly caused by app lying an inaccurate integral direction to a timestep interval. To rectify this i ssue we propose a timestep tuner that helps find a more accurate integral direct ion for a particular interval at the minimum cost. Specifically at each denoisin g step we replace the original parameterization by conditioning the network on a new timestep enforcing the sampling distribution towards the real one. Extensiv e experiments show that our plug-in design can be trained efficiently and boost the inference performance of various state-of-the-art acceleration methods espec ially when there are few denoising steps. For example when using 10 denoising st eps on LSUN Bedroom dataset we improve the FID of DDIM from 9.65 to 6.07 simply by adopting our method for a more appropriate set of timesteps. Code is availabl e at \href https://github.com/THU-LYJ-Lab/time-tuner https://github.com/THU-LYJ -Lab/time-tuner .
****
Rethinking Generalizable Face Anti-spoofing via Hierarchical Prototype-guided Di stribution Refinement in Hyperbolic Space
Chengyang Hu, Ke-Yue Zhang, Taiping Yao, Shouhong Ding, Lizhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 024, pp. 1032-1041
Generalizable face anti-spoofing (FAS) approaches have drawn growing attention d ue to their robustness for diverse presentation attacks in unseen scenarios. Mos t previous methods always utilize domain generalization (DG) frameworks via dire ctly aligning diverse source samples into a common feature space. However these methods neglect the hierarchical relations in FAS samples which may hinder the g eneralization ability by direct alignment. To address these issues we propose a novel Hierarchical Prototype-guided Distribution Refinement (HPDR) framework to learn embedding in hyperbolic space which facilitates the hierarchical relation construction. We also collaborate with prototype learning for hierarchical distr ibution refinement in hyperbolic space. In detail we propose the Hierarchical Pr ototype Learning to simultaneously guide domain alignment and improve the discri minative ability via constraining the multi-level relations between prototypes a

nd instances in hyperbolic space. Moreover we design a Prototype-oriented Classifier which further considers relations between the sample and prototypes to improve the robustness of the final decision. Extensive experiments and visualizations demonstrate the effectiveness of our method against previous competitors.
****
IIRP-Net: Iterative Inference Residual Pyramid Network for Enhanced Image Registration
Tai Ma, Suwei Zhang, Jiafeng Li, Ying Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11546-11555
Deep learning-based image registration (DLIR) methods have achieved remarkable success in deformable image registration. We observe that iterative inference can exploit the well-trained registration network to the fullest extent. In this work we propose a novel Iterative Inference Residual Pyramid Network (IIRP-Net) to enhance registration performance without any additional training costs. In IIRP-Net we construct a streamlined pyramid registration network consisting of a feature extractor and residual flow estimators (RP-Net) to achieve generalized capabilities in feature extraction and registration. Then in the inference phase IIRP-Net employs an iterative inference strategy to enhance RP-Net by iteratively reutilizing residual flow estimators from coarse to fine. The number of iterations is adaptively determined by the proposed IterStop mechanism. We conduct extensive experiments on the FLARE and Mindboggle datasets and the results verify the effectiveness of the proposed method outperforming state-of-the-art deformable image registration methods. Our code is available at https://github.com/Torbjorn1997/IIRP-Net.
****
Learning without Exact Guidance: Updating Large-scale High-resolution Land Cover Maps from Low-resolution Historical Labels
Zhuohong Li, Wei He, Jiepan Li, Fangxiao Lu, Hongyan Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27717-27727
Large-scale high-resolution (HR) land-cover mapping is a vital task to survey the Earth's surface and resolve many challenges facing humanity. However it is still a non-trivial task hindered by complex ground details various landforms and the scarcity of accurate training labels over a wide-span geographic area. In this paper we propose an efficient weakly supervised framework (Paraformer) to guide large-scale HR land-cover mapping with easy-access historical land-cover data of low resolution (LR). Specifically existing land-cover mapping approaches reveal the dominance of CNNs in preserving local ground details but still suffer from insufficient global modeling in various landforms. Therefore we design a parallel CNN-Transformer feature extractor in Paraformer consisting of a downsampling-free CNN branch and a Transformer branch to jointly capture local and global contextual information. Besides facing the spatial mismatch of training data a pseudo-label-assisted training (PLAT) module is adopted to reasonably refine LR labels for weakly supervised semantic segmentation of HR images. Experiments on two large-scale datasets demonstrate the superiority of Paraformer over other state-of-the-art methods for automatically updating HR land-cover maps from LR historical labels.
****
GenesisTex: Adapting Image Denoising Diffusion to Texture Space
Chenjian Gao, Boyan Jiang, Xinghui Li, Yingpeng Zhang, Qian Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4620-4629
We present GenesisTex a novel method for synthesizing textures for 3D geometries from text descriptions. GenesisTex adapts the pretrained image diffusion model to texture space by texture space sampling. Specifically we maintain a latent texture map for each viewpoint which is updated with predicted noise on the rendering of the corresponding viewpoint. The sampled latent texture maps are then decoded into a final texture map. During the sampling process we focus on both global and local consistency across multiple viewpoints: global consistency is achieved through the integration of style consistency mechanisms within the noise pre

diction network and low-level consistency is achieved by dynamically aligning la
tent textures. Finally we apply reference-based inpainting and img2img on denser
 views for texture refinement. Our approach overcomes the limitations of slow op
timization in distillation-based methods and instability in inpainting-based met
hods. Experiments on meshes from various sources demonstrate that our method sur
passes the baseline methods quantitatively and qualitatively.
****
TTA-EVF: Test-Time Adaptation for Event-based Video Frame Interpolation via Reli
able Pixel and Sample Estimation
Hoonhee Cho, Taewoo Kim, Yuhwan Jeong, Kuk-Jin Yoon; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25701-2
5711
Video Frame Interpolation (VFI) which aims at generating high-frame-rate videos
from low-frame-rate inputs is a highly challenging task. The emergence of bio-in
spired sensors known as event cameras which boast microsecond-level temporal res
olution has ushered in a transformative era for VFI. Nonetheless the application
 of event-based VFI techniques in domains with distinct environments from the tr
aining data can be problematic. This is mainly because event camera data distrib
ution can undergo substantial variations based on camera settings and scene cond
itions presenting challenges for effective adaptation. In this paper we propose
a test-time adaptation method for event-based VFI to address the gap between the
 source and target domains. Our approach enables sequential learning in an onlin
e manner on the target domain which only provides low-frame-rate videos. We pres
ent an approach that leverages confident pixels as pseudo ground-truths enabling
 stable and accurate online learning from low-frame-rate videos. Furthermore to
prevent overfitting during the continuous online process where the same scene is
 encountered repeatedly we propose a method of blending historical samples with
current scenes. Extensive experiments validate the effectiveness of our method b
oth in cross-domain and continuous domain shifting setups. The code is available
 at https://github.com/Chohoonhee/TTA-EVF.
****
Image-to-Image Matching via Foundation Models: A New Perspective for Open-Vocabu
lary Semantic Segmentation
Yuan Wang, Rui Sun, Naisong Luo, Yuwen Pan, Tianzhu Zhang; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3
952-3963
Open-vocabulary semantic segmentation (OVS) aims to segment images of arbitrary
categories specified by class labels or captions. However most previous best-per
forming methods whether pixel grouping methods or region recognition methods suf
fer from false matches between image features and category labels. We attribute
this to the natural gap between the textual features and visual features. In thi
s work we rethink how to mitigate false matches from the perspective of image-to
-image matching and propose a novel relation-aware intra-modal matching (RIM) fr
amework for OVS based on visual foundation models. RIM achieves robust region cl
assification by firstly constructing diverse image-modal reference features and
then matching them with region features based on relation-aware ranking distribu
tion. The proposed RIM enjoys several merits. First the intra-modal reference fe
atures are better aligned circumventing potential ambiguities that may arise in
cross-modal matching. Second the ranking-based matching process harnesses the st
ructure information implicit in the inter-class relationships making it more rob
ust than comparing individually. Extensive experiments on three benchmarks demon
strate that RIM outperforms previous state-of-the-art methods by large margins o
btaining a lead of more than 10% in mIoU on PASCAL VOC benchmark
****
BigGait: Learning Gait Representation You Want by Large Vision Models
Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, Shiqi Yu; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
200-210
Gait recognition stands as one of the most pivotal remote identification technol
ogies and progressively expands across research and industry communities. Howeve

r existing gait recognition methods heavily rely on task-specific upstream drive
n by supervised learning to provide explicit gait representations like silhouett
e sequences which inevitably introduce expensive annotation costs and potential
error accumulation. Escaping from this trend this work explores effective gait r
epresentations based on the all-purpose knowledge produced by task-agnostic Larg
e Vision Models (LVMs) and proposes a simple yet efficient gait framework termed
 BigGait. Specifically the Gait Representation Extractor (GRE) within BigGait dr
aws upon design principles from established gait representations effectively tra
nsforming all-purpose knowledge into implicit gait representations without requi
ring third-party supervision signals. Experiments on CCPG CAISA-B* and SUSTech1K
 indicate that BigGait significantly outperforms the previous methods in both wi
thin-domain and cross-domain tasks in most cases and provides a more practical p
aradigm for learning the next-generation gait representation. Finally we delve i
nto prospective challenges and promising directions in LVMs-based gait recogniti
on aiming to inspire future work in this emerging topic. The source code is avai
lable at https://github.com/ShiqiYu/OpenGait.
****
BEVNeXt: Reviving Dense BEV Frameworks for 3D Object Detection
Zhenxin Li, Shiyi Lan, Jose M. Alvarez, Zuxuan Wu; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20113-201
23
Recently the rise of query-based Transformer decoders is reshaping camera-based
3D object detection. These query-based decoders are surpassing the traditional d
ense BEV (Bird's Eye View)-based methods. However we argue that dense BEV framew
orks remain important due to their outstanding abilities in depth estimation and
 object localization depicting 3D scenes accurately and comprehensively. This pa
per aims to address the drawbacks of the existing dense BEV-based 3D object dete
ctors by introducing our proposed enhanced components including a CRF-modulated
depth estimation module enforcing object-level consistencies a long-term tempora
l aggregation module with extended receptive fields and a two-stage object decod
er combining perspective techniques with CRF-modulated depth embedding. These en
hancements lead to a "modernized" dense BEV framework dubbed BEVNeXt. On the nuS
cenes benchmark BEVNeXt outperforms both BEV-based and query-based frameworks un
der various settings achieving a state-of-the-art result of 64.2 NDS on the nuSc
enes test set.
****
SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinfor
mation Detection
Peng Qi, Zehong Yan, Wynne Hsu, Mong Li Lee; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13052-13062
Misinformation is a prevalent societal issue due to its potential high risks. Ou
t-Of-Context (OOC) misinformation where authentic images are repurposed with fal
se text is one of the easiest and most effective ways to mislead audiences. Curr
ent methods focus on assessing image-text consistency but lack convincing explan
ations for their judgments which are essential for debunking misinformation. Whi
le Multimodal Large Language Models (MLLMs) have rich knowledge and innate capab
ility for visual reasoning and explanation generation they still lack sophistica
tion in understanding and discovering the subtle cross-modal differences. In thi
s paper we introduce Sniffer a novel multimodal large language model specificall
y engineered for OOC misinformation detection and explanation. Sniffer employs t
wo-stage instruction tuning on InstructBLIP. The first stage refines the model's
 concept alignment of generic objects with news-domain entities and the second s
tage leverages OOC-specific instruction data generated by language-only GPT-4 to
 fine-tune the model's discriminatory powers. Enhanced by external tools and ret
rieval Sniffer not only detects inconsistencies between text and image but also
utilizes external knowledge for contextual verification. Our experiments show th
at Sniffer surpasses the original MLLM by over 40% and outperforms state-of-the-
art methods in detection accuracy. Sniffer also provides accurate and persuasive
 explanations as validated by quantitative and human evaluations.
****

Beyond Seen Primitive Concepts and Attribute-Object Compositional Learning
Nirat Saini, Khoi Pham, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14466-14476
Learning from seen attribute-object pairs to generalize to unseen compositions has been studied extensively in Compositional Zero-Shot Learning (CZSL). However CZSL setup is still limited to seen attributes and objects and cannot generalize to unseen concepts and their compositions. To overcome this limitation we propose a new task Open Vocabulary-Compositional Zero-shot Learning (OV-CZSL) where unseen attributes objects and unseen compositions are evaluated. To show that OV-CZSL is a challenging yet solvable problem we propose three new benchmarks based on existing datasets MIT-States C-GQA and VAW-CZSL along with new baselines and evaluation setup. We use language embeddings and external vocabulary with our novel neighborhood expansion loss to allow any method to learn semantic correlations between seen and unseen primitives.
****
Unleashing Network Potentials for Semantic Scene Completion
Fengyun Wang, Qianru Sun, Dong Zhang, Jinhui Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10314-10323
Semantic scene completion (SSC) aims to predict complete 3D voxel occupancy and semantics from a single-view RGB-D image and recent SSC methods commonly adopt multi-modal inputs. However our investigation reveals two limitations: ineffective feature learning from single modalities and overfitting to limited datasets. To address these issues this paper proposes a novel SSC framework - Adversarial Modality Modulation Network (AMMNet) - with a fresh perspective of optimizing gradient updates. The proposed AMMNet introduces two core modules: a cross-modal modulation enabling the interdependence of gradient flows between modalities and a customized adversarial training scheme leveraging dynamic gradient competition. Specifically the cross-modal modulation adaptively re-calibrates the features to better excite representation potentials from each single modality. The adversarial training employs a minimax game of evolving gradients with customized guidance to strengthen the generator's perception of visual fidelity from both geometric completeness and semantic correctness. Extensive experimental results demonstrate that AMMNet outperforms state-of-the-art SSC methods by a large margin providing a promising direction for improving the effectiveness and generalization of SSC methods.
****
HOIST-Former: Hand-held Objects Identification Segmentation and Tracking in the Wild
Supreeth Narasimhaswamy, Huy Anh Nguyen, Lihan Huang, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2351-2361
We address the challenging task of identifying segmenting and tracking hand-held objects which is crucial for applications such as human action segmentation and performance evaluation. This task is particularly challenging due to heavy occlusion rapid motion and the transitory nature of objects being hand-held where an object may be held released and subsequently picked up again. To tackle these challenges we have developed a novel transformer-based architecture called HOIST-Former. HOIST-Former is adept at spatially and temporally segmenting hands and objects by iteratively pooling features from each other ensuring that the processes of identification segmentation and tracking of hand-held objects depend on the hands' positions and their contextual appearance. We further refine HOIST-Former with a contact loss that focuses on areas where hands are in contact with objects. Moreover we also contribute an in-the-wild video dataset called HOIST which comprises 4125 videos complete with bounding boxes segmentation masks and tracking IDs for hand-held objects. Through experiments on the HOIST dataset and two additional public datasets we demonstrate the efficacy of HOIST-Former in segmenting and tracking hand-held objects.
****
Contextrast: Contextual Contrastive Learning for Semantic Segmentation

Changki Sung, Wanhee Kim, Jungho An, Wooju Lee, Hyungtae Lim, Hyun Myung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3732-3742
Despite great improvements in semantic segmentation challenges persist because of the lack of local/global contexts and the relationship between them. In this paper we propose Contextrast a contrastive learning-based semantic segmentation method that allows to capture local/global contexts and comprehend their relationships. Our proposed method comprises two parts: a) contextual contrastive learning (CCL) and b) boundary-aware negative (BANE) sampling. Contextual contrastive learning obtains local/global context from multi-scale feature aggregation and inter/intra-relationship of features for better discrimination capabilities. Meanwhile BANE sampling selects embedding features along the boundaries of incorrectly predicted regions to employ them as harder negative samples on our contrastive learning resolving segmentation issues along the boundary region by exploiting fine-grained details. We demonstrate that our Contextrast substantially enhances the performance of semantic segmentation networks outperforming state-of-the-art contrastive learning approaches on diverse public datasets e.g. Cityscapes CamVid PASCAL-C COCO-Stuff and ADE20K without an increase in computational cost during inference.
****
Learning Occupancy for Monocular 3D Object Detection
Liang Peng, Junkai Xu, Haoran Cheng, Zheng Yang, Xiaopei Wu, Wei Qian, Wenxiao Wang, Boxi Wu, Deng Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10281-10292
Monocular 3D detection is a challenging task due to the lack of accurate 3D information. Existing approaches typically rely on geometry constraints and dense depth estimates to facilitate the learning but often fail to fully exploit the benefits of three-dimensional feature extraction in frustum and 3D space. In this paper we propose OccupancyM3D a method of learning occupancy for monocular 3D detection. It directly learns occupancy in frustum and 3D space leading to more discriminative and informative 3D features and representations. Specifically by using synchronized raw sparse LiDAR point clouds we define the space status and generate voxel-based occupancy labels. We formulate occupancy prediction as a simple classification problem and design associated occupancy losses. Resulting occupancy estimates are employed to enhance original frustum/3D features. As a result experiments on KITTI and Waymo open datasets demonstrate that the proposed method achieves a new state of the art and surpasses other methods by a significant margin.
****
LAA-Net: Localized Artifact Attention Network for Quality-Agnostic and Generalizable Deepfake Detection
Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, Djamila Aouada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17395-17405
This paper introduces a novel approach for high-quality deepfake detection called Localized Artifact Attention Network (LAA-Net). Existing methods for high-quality deepfake detection are mainly based on a supervised binary classifier coupled with an implicit attention mechanism. As a result they do not generalize well to unseen manipulations. To handle this issue two main contributions are made. First an explicit attention mechanism within a multi-task learning framework is proposed. By combining heatmap-based and self-consistency attention strategies LAA-Net is forced to focus on a few small artifact-prone vulnerable regions. Second an Enhanced Feature Pyramid Network (E-FPN) is proposed as a simple and effective mechanism for spreading discriminative low-level features into the final feature output with the advantage of limiting redundancy. Experiments performed on several benchmarks show the superiority of our approach in terms of Area Under the Curve (AUC) and Average Precision (AP). The code is available at https://github.com/10Ring/LAA-Net.
****
LEAD: Learning Decomposition for Source-free Universal Domain Adaptation

Sanqing Qu, Tianpei Zou, Lianghua He, Florian Röhrbein, Alois Knoll, Guang Chen, Changjun Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23334-23343

Universal Domain Adaptation (UniDA) targets knowledge transfer in the presence of both covariate and label shifts. Recently Source-free Universal Domain Adaptation (SF-UniDA) has emerged to achieve UniDA without access to source data which tends to be more practical due to data protection policies. The main challenge lies in determining whether covariate-shifted samples belong to target-private unknown categories. Existing methods tackle this either through hand-crafted thresholding or by developing time-consuming iterative clustering strategies. In this paper we propose a new idea of LEArning Decomposition (LEAD) which decouples features into source-known and -unknown components to identify target-private data. Technically LEAD initially leverages the orthogonal decomposition analysis for feature decomposition. Then LEAD builds instance-level decision boundaries to adaptively identify target-private data. Extensive experiments across various UniDA scenarios have demonstrated the effectiveness and superiority of LEAD. Notably in the OPDA scenario on VisDA dataset LEAD outperforms GLC by 3.5% overall H-score and reduces 75% time to derive pseudo-labeling decision boundaries. Besides LEAD is also appealing in that it is complementary to most existing methods. The code is available at https://github. com/ispc-lab/LEAD

****

AUEditNet: Dual-Branch Facial Action Unit Intensity Manipulation with Implicit Disentanglement

Shiwei Jin, Zhen Wang, Lei Wang, Peng Liu, Ning Bi, Truong Nguyen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2104-2113

Facial action unit (AU) intensity plays a pivotal role in quantifying fine-grained expression behaviors which is an effective condition for facial expression manipulation. However publicly available datasets containing intensity annotations for multiple AUs remain severely limited often featuring a restricted number of subjects. This limitation places challenges to the AU intensity manipulation in images due to disentanglement issues leading researchers to resort to other large datasets with pretrained AU intensity estimators for pseudo labels. In addressing this constraint and fully leveraging manual annotations of AU intensities for precise manipulation we introduce AUEditNet. Our proposed model achieves impressive intensity manipulation across 12 AUs trained effectively with only 18 subjects. Utilizing a dual-branch architecture our approach achieves comprehensive disentanglement of facial attributes and identity without necessitating additional loss functions or implementing with large batch sizes. This approach offers a potential solution to achieve desired facial attribute editing despite the dataset's limited subject count. Our experiments demonstrate AUEditNet's superior accuracy in editing AU intensities affirming its capability in disentangling facial attributes and identity within a limited subject pool. AUEditNet allows conditioning by either intensity values or target images eliminating the need for constructing AU combinations for specific facial expression synthesis. Moreover AU intensity estimation as a downstream task validates the consistency between real and edited images confirming the effectiveness of our proposed AU intensity manipulation method.

****

BodyMAP - Jointly Predicting Body Mesh and 3D Applied Pressure Map for People in Bed

Abhishek Tandon, Anujraaj Goyal, Henry M. Clever, Zackory Erickson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2480-2489

Accurately predicting the 3D human posture and the pressure exerted on the body for people resting in bed visualized as a body mesh (3D pose & shape) with a 3D pressure map holds significant promise for healthcare applications particularly in the prevention of pressure ulcers. Current methods focus on singular facets of the problem---predicting only 2D/3D poses generating 2D pressure images predicting pressure only for certain body regions instead of the full body or forming

indirect approximations to the 3D pressure map. In contrast we introduce BodyMAP
 which jointly predicts the human body mesh and 3D applied pressure map across t
he entire human body. Our network leverages multiple visual modalities incorpora
ting both a depth image of a person in bed and its corresponding 2D pressure ima
ge acquired from a pressure-sensing mattress. The 3D pressure map is represented
 as a pressure value at each mesh vertex and thus allows for precise localizatio
n of high-pressure regions on the body. Additionally we present BodyMAP-WS a new
 formulation of pressure prediction in which we implicitly learn pressure in 3D
by aligning sensed 2D pressure images with a differentiable 2D projection of the
 predicted 3D pressure maps. In evaluations with real-world human data our metho
d outperforms the current state-of-the-art technique by 25% on both body mesh an
d 3D applied pressure map prediction tasks for people in bed.
****
OneLLM: One Framework to Align All Modalities with Language
Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin,
Yu Qiao, Peng Gao, Xiangyu Yue; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2024, pp. 26584-26595
Multimodal large language models (MLLMs) have gained significant attention due t
o their strong multimodal understanding capability. However existing works rely
heavily on modality-specific encoders which usually differ in architecture and a
re limited to common modalities. In this paper we present OneLLM an MLLM that al
igns eight modalities to language using a unified framework. We achieve this thr
ough a unified multimodal encoder and a progressive multimodal alignment pipelin
e. In detail we first train an image projection module to connect a vision encod
er with LLM. Then we build a universal projection module (UPM) by mixing multipl
e image projection modules and dynamic routing. Finally we progressively align m
ore modalities to LLM with the UPM. To fully leverage the potential of OneLLM in
 following instructions we also curated a comprehensive multimodal instruction d
ataset including 2M items from image audio video point cloud depth/normal map IM
U and fMRI brain activity. OneLLM is evaluated on 25 diverse benchmarks encompas
sing tasks such as multimodal captioning question answering and reasoning where
it delivers excellent performance. Code data model and online demo are available
 at https://github.com/csuhan/OneLLM
****
PAD: Patch-Agnostic Defense against Adversarial Patch Attacks
Lihua Jing, Rui Wang, Wenqi Ren, Xin Dong, Cong Zou; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24472-2
4481
Adversarial patch attacks present a significant threat to real-world object dete
ctors due to their practical feasibility. Existing defense methods which rely on
 attack data or prior knowledge struggle to effectively address a wide range of
adversarial patches. In this paper we show two inherent characteristics of adver
sarial patches semantic independence and spatial heterogeneity independent of th
eir appearance shape size quantity and location. Semantic independence indicates
 that adversarial patches operate autonomously within their semantic context whi
le spatial heterogeneity manifests as distinct image quality of the patch area t
hat differs from original clean image due to the independent generation process.
 Based on these observations we propose PAD a novel adversarial patch localizati
on and removal method that does not require prior knowledge or additional traini
ng. PAD offers patch-agnostic defense against various adversarial patches compat
ible with any pre-trained object detectors. Our comprehensive digital and physic
al experiments involving diverse patch types such as localized noise printable a
nd naturalistic patches exhibit notable improvements over state-of-the-art works
. Our code is available at https://github.com/Lihua-Jing/PAD.
****
MULAN: A Multi Layer Annotated Dataset for Controllable Text-to-Image Generation
Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Ge
rasimos Lampouras, Ignacio Iacobacci, Sarah Parisot; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22413-2
2422

Text-to-image generation has achieved astonishing results yet precise spatial co ntrollability and prompt fidelity remain highly challenging. This limitation is typically addressed through cumbersome prompt engineering scene layout condition ing or image editing techniques which often require hand drawn masks. Nonetheles s pre-existing works struggle to take advantage of the natural instance-level co mpositionality of scenes due to the typically flat nature of rasterized RGB outp ut images. Towards addressing this challenge we introduce MuLAn: a novel dataset comprising over 44K MUlti-Layer ANnotations of RGB images as multi-layer instan ce-wise RGBA decompositions and over 100K instance images. To build MuLAn we dev eloped a training free pipeline which decomposes a monocular RGB image into a st ack of RGBA layers comprising of background and isolated instances. We achieve t his through the use of pretrained general-purpose models and by developing three modules: image decomposition for instance discovery and extraction instance com pletion to reconstruct occluded areas and image re-assembly. We use our pipeline to create MuLAn-COCO and MuLAn-LAION datasets which contain a variety of image decompositions in terms of style composition and complexity. With MuLAn we provi de the first photorealistic resource providing instance decomposition and occlus ion information for high quality images opening up new avenues for text-to-image generative AI research. With this we aim to encourage the development of novel generation and editing technology in particular layer-wise solutions. MuLAn data resources are available at https://MuLAn-dataset.github.io/
****
Rotation-Agnostic Image Representation Learning for Digital Pathology
Saghir Alfasly, Abubakr Shafique, Peyman Nejat, Jibran Khan, Areej Alsaafin, Gha zal Alabtah, H.R. Tizhoosh; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2024, pp. 11683-11693
This paper addresses complex challenges in histopathological image analysis thro ugh three key contributions. Firstly it introduces a fast patch selection method FPS for whole-slide image (WSI) analysis significantly reducing computational c ost while maintaining accuracy. Secondly it presents PathDino a lightweight hist opathology feature extractor with a minimal configuration of five Transformer bl ocks and only ? 9 million parameters markedly fewer than alternatives. Thirdly i t introduces a rotation-agnostic representation learning paradigm using self-sup ervised learning effectively mitigating overfitting. We also show that our compa ct model outperforms existing state-of-the-art histopathology-specific vision tr ansformers on 12 diverse datasets including both internal datasets spanning four sites (breast liver skin and colorectal) and seven public datasets (PANDA CAMEL YON16 BRACS DigestPath Kather PanNuke and WSSS4LUAD). Notably even with a traini ng dataset of ? 6 million histopathology patches from The Cancer Genome Atlas (T CGA) our approach demonstrates an average 8.5% improvement in patch-level majori ty vote performance. These contributions provide a robust framework for enhancin g image analysis in digital pathology rigorously validated through extensive eva luation.
****
Unbiased Faster R-CNN for Single-source Domain Generalized Object Detection
Yajing Liu, Shijun Zhou, Xiyao Liu, Chunhui Hao, Baojie Fan, Jiandong Tian; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2024, pp. 28838-28847
Single-source domain generalization (SDG) for object detection is a challenging yet essential task as the distribution bias of the unseen domain degrades the al gorithm performance significantly. However existing methods attempt to extract d omain-invariant features neglecting that the biased data leads the network to le arn biased features that are non-causal and poorly generalizable. To this end we propose an Unbiased Faster R-CNN (UFR) for generalizable feature learning. Spec ifically we formulate SDG in object detection from a causal perspective and cons truct a Structural Causal Model (SCM) to analyze the data bias and feature bias in the task which are caused by scene confounders and object attribute confounde rs. Based on the SCM we design a Global-Local Transformation module for data aug mentation which effectively simulates domain diversity and mitigates the data bi as. Additionally we introduce a Causal Attention Learning module that incorporat

es a designed attention invariance loss to learn image-level features that are robust to scene confounders. Moreover we develop a Causal Prototype Learning module with an explicit instance constraint and an implicit prototype constraint which further alleviates the negative impact of object attribute confounders. Experimental results on five scenes demonstrate the prominent generalization ability of our method with an improvement of 3.9% mAP on the Night-Clear scene.
****
Super-Resolution Reconstruction from Bayer-Pattern Spike Streams
Yanchen Dong, Ruiqin Xiong, Jian Zhang, Zhaofei Yu, Xiaopeng Fan, Shuyuan Zhu, Tiejun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24871-24880
Spike camera is a neuromorphic vision sensor that can capture highly dynamic scenes by generating a continuous stream of binary spikes to represent the arrival of photons at very high temporal resolution. Equipped with Bayer color filter array (CFA) color spike camera (CSC) has been invented to capture color information. Although spike camera has already demonstrated great potential for high-speed imaging its spatial resolution is limited compared with conventional digital cameras. This paper proposes a Color Spike Camera Super-Resolution (CSCSR) network to super-resolve higher-resolution color images from spike camera streams with Bayer CFA. To be specific we first propose a representation for Bayer-pattern spike streams exploring local temporal information with global perception to represent the binary data. Then we exploit the CFA layout and sub-pixel level motion to collect temporal pixels for the spatial super-resolution of each color channel. In particular a residual-based module for feature refinement is developed to reduce the impact of motion estimation errors. Considering color correlation we jointly utilize the multi-stage temporal-pixel features of color channels to reconstruct the high-resolution color image. Experimental results demonstrate that the proposed scheme can reconstruct satisfactory color images with both high temporal and spatial resolution from low-resolution Bayer-pattern spike streams. The source codes are available at https://github.com/csycdong/CSCSR.
****
EASE-DETR: Easing the Competition among Object Queries
Yulu Gao, Yifan Sun, Xudong Ding, Chuyang Zhao, Si Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17282-17291
This paper views the DETR's non-duplicate detection ability as a competition result among object queries. Around each object there are usually multiple queries within which only a single one can win the chance to become the final detection. Such a competition is hard: while some competing queries initially have very close prediction scores their leading query has to dramatically enlarge its score superiority after several decoder layers. To help the leading query stands out this paper proposes EASE-DETR which eases the competition by introducing bias that favours the leading one. EASE-DETR is very simple: in every intermediate decoder layer we identify the "leading / trailing" relationship between any two queries and encode this binary relationship into the following decoder layer to amplify the superiority of the leading one. More concretely the leading query is to be protected from mutual query suppression in the self-attention layer and encouraged to absorb more object features in the cross-attention layer therefore accelerating to win. Experimental results show that EASE-DETR brings consistent and remarkable improvement to various DETRs.
****
KPConvX: Modernizing Kernel Point Convolution with Kernel Attention
Hugues Thomas, Yao-Hung Hubert Tsai, Timothy D. Barfoot, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5525-5535
In the field of deep point cloud understanding KPConv is a unique architecture that uses kernel points to locate convolutional weights in space instead of relying on Multi-Layer Perceptron (MLP) encodings. While it initially achieved success it has since been surpassed by recent MLP networks that employ updated designs and training strategies. Building upon the kernel point principle we present tw

o novel designs: KPConvD (depthwise KPConv) a lighter design that enables the us
e of deeper architectures and KPConvX an innovative design that scales the depth
wise convolutional weights of KPConvD with kernel attention values. Using KPConv
X with a modern architecture and training strategy we are able to outperform cur
rent state-of-the-art approaches on the ScanObjectNN Scannetv2 and S3DIS dataset
s. We validate our design choices through ablation studies and release our code
and models.
****
Clockwork Diffusion: Efficient Generation With Model-Step Distillation
Amirhossein Habibian, Amir Ghodrati, Noor Fathima, Guillaume Sautiere, Risheek G
arrepalli, Fatih Porikli, Jens Petersen; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8352-8361
This work aims to improve the efficiency of text-to-image diffusion models. Whil
e diffusion models use computationally expensive UNet-based denoising operations
 in every generation step we identify that not all operations are equally releva
nt for the final output quality. In particular we observe that UNet layers opera
ting on high-res feature maps are relatively sensitive to small perturbations. I
n contrast low-res feature maps influence the semantic layout of the final image
 and can often be perturbed with no noticeable change in the output. Based on th
is observation we propose Clockwork Diffusion a method that periodically reuses
computation from preceding denoising steps to approximate low-res feature maps a
t one or more subsequent steps. For multiple base- lines and for both text-to-im
age generation and image editing we demonstrate that Clockwork leads to comparab
le or improved perceptual scores with drastically reduced computational complexi
ty. As an example for Stable Diffusion v1.5 with 8 DPM++ steps we save 32% of FL
OPs with negligible FID and CLIP change. We re- lease code at https://github.com
/Qualcomm-AI-research/clockwork-diffusion
****
Pick-or-Mix: Dynamic Channel Sampling for ConvNets
Ashish Kumar, Daneul Kim, Jaesik Park, Laxmidhar Behera; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 587
3-5882
Channel pruning approaches for convolutional neural networks (ConvNets) deactiva
te the channels statically or dynamically and require special implementation. In
 addition channel squeezing in representative ConvNets is carried out via 1 x 1
convolutions which dominates a large portion of computations and network paramet
ers. Given these challenges we propose an effective multi-purpose module for dyn
amic channel sampling namely Pick-or-Mix (PiX) which does not require special im
plementation. PiX divides a set of channels into subsets and then picks from the
m where the picking decision is dynamically made per each pixel based on the inp
ut activations. We plug PiX into prominent ConvNet architectures and verify its
multi-purpose utilities. After replacing 1 x 1 channel squeezing layers in ResNe
t with PiX the network becomes 25% faster without losing accuracy. We show that
PiX allows ConvNets to learn better data representation than widely adopted appr
oaches to enhance networks' representation power (e.g. SE CBAM AFF SKNet and DWP
). We also show that PiX achieves state-of-the-art performance on network downsc
aling and dynamic channel pruning applications.
****
Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-
to-Image Generation
Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, Jindong Gu; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 12006-12016
Diffusion-based models have gained significant popularity for text-to-image gene
ration due to their exceptional image-generation capabilities. A risk with these
 models is the potential generation of inappropriate content such as biased or h
armful images. However the underlying reasons for generating such undesired cont
ent from the perspective of the diffusion model's internal representation remain
 unclear. Previous work interprets vectors in an interpretable latent space of d
iffusion models as semantic concepts. However existing approaches cannot discove

r directions for arbitrary concepts such as those related to inappropriate conce
pts. In this work we propose a novel self-supervised approach to find interpreta
ble latent directions for a given concept. With the discovered vectors we furthe
r propose a simple approach to mitigate inappropriate generation. Extensive expe
riments have been conducted to verify the effectiveness of our mitigation approa
ch namely for fair generation safe generation and responsible text-enhancing gen
eration. Project page: https://interpretdiffusion.github.io.
****
HiLo: Detailed and Robust 3D Clothed Human Reconstruction with High-and Low-Freq
uency Information of Parametric Models
Yifan Yang, Dong Liu, Shuhai Zhang, Zeshuai Deng, Zixiong Huang, Mingkui Tan; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 10671-10681
Reconstructing 3D clothed human involves creating a detailed geometry of individ
uals in clothing with applications ranging from virtual try-on movies to games.
To enable practical and widespread applications recent advances propose to gener
ate a clothed human from an RGB image. However they struggle to reconstruct deta
iled and robust avatars simultaneously. We empirically find that the high-freque
ncy (HF) and low-frequency (LF) information from a parametric model has the pote
ntial to enhance geometry details and improve robustness to noise respectively.
Based on this we propose HiLo namely clothed human reconstruction with high- and
 low-frequency information which contains two components. 1) To recover detailed
 geometry using HF information we propose a progressive HF Signed Distance Funct
ion to enhance the detailed 3D geometry of a clothed human. We analyze that our
progressive learning manner alleviates large gradients that hinder model converg
ence. 2) To achieve robust reconstruction against inaccurate estimation of the p
arametric model by using LF information we propose a spatial interaction implici
t function. This function effectively exploits the complementary spatial informa
tion from a low-resolution voxel grid of the parametric model. Experimental resu
lts demonstrate that HiLo outperforms the state-of-the-art methods by 10.43% and
 9.54% in terms of Chamfer distance on the Thuman2.0 and CAPE datasets respectiv
ely. Additionally HiLo demonstrates robustness to noise from the parametric mode
l challenging poses and various clothing styles.
****
Promptable Behaviors: Personalizing Multi-Objective Rewards from Human Preferenc
es
Minyoung Hwang, Luca Weihs, Chanwoo Park, Kimin Lee, Aniruddha Kembhavi, Kiana E
hsani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2024, pp. 16216-16226
Customizing robotic behaviors to be aligned with diverse human preferences is an
 underexplored challenge in the field of embodied AI. In this paper we present P
romptable Behaviors a novel framework that facilitates efficient personalization
 of robotic agents to diverse human preferences in complex environments. We use
multi-objective reinforcement learning to train a single policy adaptable to a b
road spectrum of preferences. We introduce three distinct methods to infer human
 preferences by leveraging different types of interactions: (1) human demonstrat
ions (2) preference feedback on trajectory comparisons and (3) language instruct
ions. We evaluate the proposed method in personalized object-goal navigation and
 flee navigation tasks in ProcTHOR and RoboTHOR demonstrating the ability to pro
mpt agent behaviors to satisfy human preferences in various scenarios.
****
Stationary Representations: Optimally Approximating Compatibility and Implicatio
ns for Improved Model Replacements
Niccolò Biondi, Federico Pernici, Simone Ricci, Alberto Del Bimbo; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 28793-28804
Learning compatible representations enables the interchangeable use of semantic
features as models are updated over time. This is particularly relevant in searc
h and retrieval systems where it is crucial to avoid reprocessing of the gallery
 images with the updated model. While recent research has shown promising empiri

cal evidence there is still a lack of comprehensive theoretical understanding ab
out learning compatible representations. In this paper we demonstrate that the s
tationary representations learned by the d-Simplex fixed classifier optimally ap
proximate compatibility representation according to the two inequality constrain
ts of its formal definition. This not only establishes a solid foundation for fu
ture works in this line of research but also presents implications that can be e
xploited in practical learning scenarios. An exemplary application is the now-st
andard practice of downloading and fine-tuning new pre-trained models. Specifica
lly we show the strengths and critical issues of stationary representations in t
he case in which a model undergoing sequential fine-tuning is asynchronously rep
laced by downloading a better-performing model pre-trained elsewhere. Such a rep
resentation enables seamless delivery of retrieval service (i.e. no reprocessing
 of gallery images) and offers improved performance without operational disrupti
ons during model replacement. Code available at: https://github.com/miccunifi/ia
mcl2r.
****
Towards Calibrated Multi-label Deep Neural Networks
Jiacheng Cheng, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2024, pp. 27589-27599
The problem of calibrating deep neural networks (DNNs) for multi-label learning
is considered. It is well-known that DNNs trained by cross-entropy for single-la
bel or one-hot classification are poorly calibrated. Many calibration techniques
 have been proposed to address the problem. However little attention has been pa
id to the calibration of multi-label DNNs. In this literature the focus has been
 on improving labeling accuracy in the face of severe dataset unbalance. This is
 addressed by the introduction of asymmetric losses which have became very popul
ar. However these losses do not induce well calibrated classifiers. In this work
 we first provide a theoretical explanation for this poor calibration performanc
e by showing that these loses losses lack the strictly proper property a necessa
ry condition for accurate probability estimation. To overcome this problem we pr
opose a new Strictly Proper Asymmetric (SPA) loss. This is complemented by a Lab
el Pair Regularizer (LPR) that increases the number of calibration constraints i
ntroduced per training example. The effectiveness of both contributions is valid
ated by extensive experiments on various multi-label datasets. The resulting tra
ining method is shown to significantly decrease the calibration error while main
taining state-of-the-art accuracy.
****
SceneTex: High-Quality Texture Synthesis for Indoor Scenes via Diffusion Priors
Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, Matthias Nießner;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2024, pp. 21081-21091
We propose SceneTex a novel method for effectively generating high-quality and s
tyle-consistent textures for indoor scenes using depth-to-image diffusion priors
. Unlike previous methods that either iteratively warp 2D views onto a mesh surf
ace or distillate diffusion latent features without accurate geometric and style
 cues SceneTex formulates the texture synthesis task as an optimization problem
in the RGB space where style and geometry consistency are properly reflected. At
 its core SceneTex proposes a multiresolution texture field to implicitly encode
 the mesh appearance. We optimize the target texture via a score-distillation-ba
sed objective function in respective RGB renderings. To further secure the style
 consistency across views we introduce a cross-attention decoder to predict the
RGB values by cross-attending to the pre-sampled reference locations in each ins
tance. SceneTex enables various and accurate texture synthesis for 3D-FRONT scen
es demonstrating significant improvements in visual quality and prompt fidelity
over the prior texture generation methods.
****
Neural Underwater Scene Representation
Yunkai Tang, Chengxuan Zhu, Renjie Wan, Chao Xu, Boxin Shi; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
11780-11789

Among the numerous efforts towards digitally recovering the physical world Neural Radiance Fields (NeRFs) have proved effective in most cases. However underwater scene introduces unique challenges due to the absorbing water medium the local change in lighting and the dynamic contents in the scene. We aim at developing a neural underwater scene representation for these challenges modeling the complex process of attenuation unstable in-scattering and moving objects during light transport. The proposed method can reconstruct the scenes from both established datasets and in-the-wild videos with outstanding fidelity.
****

Progress-Aware Online Action Segmentation for Egocentric Procedural Task Videos

Yuhan Shen, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18186-18197

We address the problem of online action segmentation for egocentric procedural task videos. While previous studies have mostly focused on offline action segmentation where entire videos are available for both training and inference the transition to online action segmentation is crucial for practical applications like AR/VR task assistants. Notably applying an offline-trained model directly to online inference results in a significant performance drop due to the inconsistency between training and inference. We propose an online action segmentation framework by first modifying existing architectures to make them causal. Second we develop a novel action progress prediction module to dynamically estimate the progress of ongoing actions and using them to refine the predictions of causal action segmentation. Third we propose to learn task graphs from training videos and leverage them to obtain smooth and procedure-consistent segmentations. With the combination of progress and task graph with casual action segmentation our framework effectively addresses prediction uncertainty and oversegmentation in online action segmentation and achieves significant improvement on three egocentric datasets.
****

TUMTraf V2X Cooperative Perception Dataset

Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, Alois C. Knoll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22668-22677

Cooperative perception offers several benefits for enhancing the capabilities of autonomous vehicles and improving road safety. Using roadside sensors in addition to onboard sensors increases reliability and extends the sensor range. External sensors offer higher situational awareness for automated vehicles and prevent occlusions. We propose CoopDet3D a cooperative multi-modal fusion model and TUMTraf-V2X a perception dataset for the cooperative 3D object detection and tracking task. Our dataset contains 2000 labeled point clouds and 5000 labeled images from five roadside and four onboard sensors. It includes 30k 3D boxes with track IDs and precise GPS and IMU data. We labeled nine categories and covered occlusion scenarios with challenging driving maneuvers like traffic violations near-miss events overtaking and U-turns. Through multiple experiments we show that our CoopDet3D camera-LiDAR fusion model achieves an increase of +14.36 3D mAP compared to a vehicle camera-LiDAR fusion model. Finally we make our dataset model labeling tool and devkit publicly available on our website: https://tum-traffic-dataset.github.io/tumtraf-v2x.
****

Constrained Layout Generation with Factor Graphs

Mohammed Haroon Dupty, Yanfei Dong, Sicong Leng, Guoji Fu, Yong Liang Goh, Wei Lu, Wee Sun Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12851-12860

This paper addresses the challenge of object-centric layout generation under spatial constraints seen in multiple domains including floorplan design process. The design process typically involves specifying a set of spatial constraints that include object attributes like size and inter-object relations such as relative positioning. Existing works which typically represent objects as single nodes lack the granularity to accurately model complex interactions between objects. For instance often only certain parts of an object like a room's right wall intera

ct with adjacent objects. To address this gap we introduce a factor graph based approach with four latent variable nodes for each room and a factor node for each constraint. The factor nodes represent dependencies among the variables to which they are connected effectively capturing constraints that are potentially of a higher order. We then develop message-passing on the bipartite graph forming a factor graph neural network that is trained to produce a floorplan that aligns with the desired requirements. Our approach is simple and generates layouts faithful to the user requirements demonstrated by a large improvement in IOU scores over existing methods. Additionally our approach being inferential and accurate is well-suited to the practical human-in-the-loop design process where specifications evolve iteratively offering a practical and powerful tool for AI-guided design.
****
SLICE: Stabilized LIME for Consistent Explanations for Image Classification
Revoti Prasad Bora, Philipp Terhörst, Raymond Veldhuis, Raghavendra Ramachandra, Kiran Raja; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10988-10996
Local Interpretable Model-agnostic Explanations (LIME) - a widely used post-ad-hoc model agnostic explainable AI (XAI) technique. It works by training a simple transparent (surrogate) model using random samples drawn around the neighborhood of the instance (image) to be explained (IE). Explanations are then extracted for a black-box model and a given IE using the surrogate model. However the explanations of LIME suffer from inconsistency across different runs for the same model and the same IE. We identify two main types of inconsistencies: variance in the sign and importance ranks of the segments (superpixels). These factors hinder LIME from obtaining consistent explanations. We analyze these inconsistencies and propose a new method Stabilized LIME for Consistent Explanations (SLICE). The proposed method handles the stabilization problem in two aspects: using a novel feature selection technique to eliminate spurious superpixels and an adaptive perturbation technique to generate perturbed images in the neighborhood of IE. Our results demonstrate that the explanations from SLICE exhibit significantly better consistency and fidelity than LIME (and its variant BayLime).
****
Anomaly Heterogeneity Learning for Open-set Supervised Anomaly Detection
Jiawen Zhu, Choubo Ding, Yu Tian, Guansong Pang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17616-17626
Open-set supervised anomaly detection (OSAD) - a recently emerging anomaly detection area - aims at utilizing a few samples of anomaly classes seen during training to detect unseen anomalies (i.e. samples from open-set anomaly classes) while effectively identifying the seen anomalies. Benefiting from the prior knowledge illustrated by the seen anomalies current OSAD methods can often largely reduce false positive errors. However these methods are trained in a closed-set setting and treat the anomaly examples as from a homogeneous distribution rendering them less effective in generalizing to unseen anomalies that can be drawn from any distribution. This paper proposes to learn heterogeneous anomaly distributions using the limited anomaly examples to address this issue. To this end we introduce a novel approach namely Anomaly Heterogeneity Learning (AHL) that simulates a diverse set of heterogeneous anomaly distributions and then utilizes them to learn a unified heterogeneous abnormality model in surrogate open-set environments. Further AHL is a generic framework that existing OSAD models can plug and play for enhancing their abnormality modeling. Extensive experiments on nine real-world anomaly detection datasets show that AHL can 1) substantially enhance different state-of-the-art OSAD models in detecting seen and unseen anomalies and 2) effectively generalize to unseen anomalies in new domains. Code is available at https://github.com/mala-lab/AHL.
****
SPECAT: SPatial-spEctral Cumulative-Attention Transformer for High-Resolution Hyperspectral Image Reconstruction
Zhiyang Yao, Shuyang Liu, Xiaoyun Yuan, Lu Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25368-25377

Compressive spectral image reconstruction is a critical method for acquiring ima
ges with high spatial and spectral resolution. Current advanced methods which in
volve designing deeper networks or adding more self-attention modules are limite
d by the scope of attention modules and the irrelevance of attentions across dif
ferent dimensions. This leads to difficulties in capturing non-local mutation fe
atures in the spatial-spectral domain and results in a significant parameter inc
rease but only limited performance improvement. To address these issues we propo
se SPECAT a SPatial-spEctral Cumulative-Attention Transformer designed for high-
resolution hyperspectral image reconstruction. SPECAT utilizes Cumulative-Attent
ion Blocks (CABs) within an efficient hierarchical framework to extract features
 from non-local spatial-spectral details. Furthermore it employs a projection-ob
ject Dual-domain Loss Function (DLF) to integrate the optical path constraint a
physical aspect often overlooked in current methodologies. Ultimately SPECAT not
 only significantly enhances the reconstruction quality of spectral details but
also breaks through the bottleneck of mutual restriction between the cost and ac
curacy in existing algorithms. Our experimental results demonstrate the superior
ity of SPECAT achieving 40.3 dB in hyperspectral reconstruction benchmarks outpe
rforming the state-of-the-art (SOTA) algorithms by 1.2 dB while using only 5% of
 the network parameters and 10% of the computational cost. The code is available
 at https://github.com/THU-luvision/SPECAT.
****
Attentive Illumination Decomposition Model for Multi-Illuminant White Balancing
Dongyoung Kim, Jinwoo Kim, Junsang Yu, Seon Joo Kim; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25512-2
5521
White balance (WB) algorithms in many commercial cameras assume single and unifo
rm illumination leading to undesirable results when multiple lighting sources wi
th different chromaticities exist in the scene. Prior research on multi-illumina
nt WB typically predicts illumination at the pixel level without fully grasping
the scene's actual lighting conditions including the number and color of light s
ources. This often results in unnatural outcomes lacking in overall consistency.
 To handle this problem we present a deep white balancing model that leverages t
he slot attention where each slot is in charge of representing individual illumi
nants. This design enables the model to generate chromaticities and weight maps
for individual illuminants which are then fused to compose the final illuminatio
n map. Furthermore we propose the centroid-matching loss which regulates the act
ivation of each slot based on the color range thereby enhancing the model to sep
arate illumination more effectively. Our method achieves the state-of-the-art pe
rformance on both single- and multi-illuminant WB benchmarks and also offers add
itional information such as the number of illuminants in the scene and their chr
omaticity. This capability allows for illumination editing an application not fe
asible with prior methods.
****
Efficient Stitchable Task Adaptation
Haoyu He, Zizheng Pan, Jing Liu, Jianfei Cai, Bohan Zhuang; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
28555-28565
The paradigm of pre-training and fine-tuning has laid the foundation for deployi
ng deep learning models. However most fine-tuning methods are designed to meet a
 specific resource budget. Recently considering diverse deployment scenarios wit
h various resource budgets SN-Net is introduced to quickly obtain numerous new n
etworks (stitches) from the pre-trained models (anchors) in a model family via m
odel stitching. Although promising SN-Net confronts new challenges when adapting
 it to new target domains including huge memory and storage requirements and a l
ong and sub-optimal multistage adaptation process. In this work we present a nov
el framework Efficient Stitchable Task Adaptation (ESTA) to efficiently produce
a palette of fine-tuned models that adhere to diverse resource constraints. Spec
ifically we first tailor parameter-efficient fine-tuning to share low-rank updat
es among the stitches while maintaining independent bias terms. In this way we l
argely reduce fine-tuning memory burdens and mitigate the interference among sti

tches that arises in task adaptation. Furthermore we streamline a simple yet eff ective one-stage deployment pipeline which estimates the important stitches to d eploy with training-time gradient statistics. By assigning higher sampling proba bilities to important stitches we also get a boosted Pareto frontier. Extensive experiments on 25 downstream visual recognition tasks demonstrate that our ESTA is capable of generating stitches with smooth accuracy-efficiency trade-offs and surpasses the direct SN-Net adaptation by remarkable margins with significantly lower training time and fewer trainable parameters. Furthermore we demonstrate the flexibility and scalability of our ESTA framework by stitching LLMs from LLa MA family obtaining chatbot stitches of assorted sizes.
****
Image Processing GNN: Breaking Rigidity in Super-Resolution
Yuchuan Tian, Hanting Chen, Chao Xu, Yunhe Wang; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24108-24117
Super-Resolution (SR) reconstructs high-resolution images from low-resolution on es. CNNs and window-attention methods are two major categories of canonical SR m odels. However these measures are rigid: in both operations each pixel gathers t he same number of neighboring pixels hindering their effectiveness in SR tasks. Alternatively we leverage the flexibility of graphs and propose the Image Proces sing GNN (IPG) model to break the rigidity that dominates previous SR methods. F irstly SR is unbalanced in that most reconstruction efforts are concentrated to a small proportion of detail-rich image parts. Hence we leverage degree flexibil ity by assigning higher node degrees to detail-rich image nodes. Then in order t o construct graphs for SR-effective aggregation we treat images as pixel node se ts rather than patch nodes. Lastly we hold that both local and global informatio n are crucial for SR performance. In the hope of gathering pixel information fro m both local and global scales efficiently via flexible graphs we search node co nnections within nearby regions to construct local graphs; and find connections within a strided sampling space of the whole image for global graphs. The flexib ility of graphs boosts the SR performance of the IPG model. Experiment results o n various datasets demonstrates that the proposed IPG outperforms State-of-the-A rt baselines. Codes are available at https://github.com/huawei-noah/Efficient-Co mputing/tree/master/LowLevel/IPG.
****
Revisiting Counterfactual Problems in Referring Expression Comprehension
Zhihan Yu, Ruifan Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13438-13448
Traditional referring expression comprehension (REC) aims to locate the target r eferent in an image guided by a text query. Several previous methods have studie d on the Counterfactual problem in REC (C-REC) where the objects for a given que ry cannot be found in the image. However these methods focus on the overall imag e-text or specific attribute mismatch only. In this paper we address the C-REC p roblem from a deep perspective of fine-grained attributes. To this aim we first propose a fine-grained counterfactual sample generation method to construct C-RE C datasets. Specifically we leverage pre-trained language model such as BERT to modify the attribute words in the queries obtaining the corresponding counterfac tual samples. Furthermore we propose a C-REC framework. We first adopt three enc oders to extract image text and attribute features. Then our dual-branch attenti ve fusion module fuses these cross-modal features with two branches by an attent ion mechanism. At last two prediction heads generate a bounding box and a counte rfactual label respectively. In addition we incorporate contrastive learning wit h the generated counterfactual samples as negatives to enhance the counterfactua l perception. Extensive experiments show that our framework achieves promising p erformance on both public REC datasets RefCOCO/+/g and our constructed C-REC dat asets C-RefCOCO/+/g. The code and data are available at https://github.com/Glaci er0012/CREC.
****
DyBluRF: Dynamic Neural Radiance Fields from Blurry Monocular Video
Huiqiang Sun, Xingyi Li, Liao Shen, Xinyi Ye, Ke Xian, Zhiguo Cao; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202

4, pp. 7517-7527
Recent advancements in dynamic neural radiance field methods have yielded remark
able outcomes. However these approaches rely on the assumption of sharp input im
ages. When faced with motion blur existing dynamic NeRF methods often struggle t
o generate high-quality novel views. In this paper we propose DyBluRF a dynamic
radiance field approach that synthesizes sharp novel views from a monocular vide
o affected by motion blur. To account for motion blur in input images we simulta
neously capture the camera trajectory and object Discrete Cosine Transform (DCT)
 trajectories within the scene. Additionally we employ a global cross-time rende
ring approach to ensure consistent temporal coherence across the entire scene. W
e curate a dataset comprising diverse dynamic scenes that are specifically tailo
red for our task. Experimental results on our dataset demonstrate that our metho
d outperforms existing approaches in generating sharp novel views from motion-bl
urred inputs while maintaining spatial-temporal consistency of the scene.
****
Compressed 3D Gaussian Splatting for Accelerated Novel View Synthesis
Simon Niedermayr, Josef Stumpfegger, Rüdiger Westermann; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 103
49-10358
Recently high-fidelity scene reconstruction with an optimized 3D Gaussian splat
representation has been introduced for novel view synthesis from sparse image se
ts. Making such representations suitable for applications like network streaming
 and rendering on low-power devices requires significantly reduced memory consum
ption as well as improved rendering efficiency. We propose a compressed 3D Gauss
ian splat representation that utilizes sensitivity-aware vector clustering with
quantization-aware training to compress directional colors and Gaussian paramete
rs. The learned codebooks have low bitrates and achieve a compression rate of up
 to 31 on real-world scenes with only minimal degradation of visual quality. We
demonstrate that the compressed splat representation can be efficiently rendered
 with hardware rasterization on lightweight GPUs at up to 4 higher framerates th
an reported via an optimized GPU compute pipeline. Extensive experiments across
multiple datasets demonstrate the robustness and rendering speed of the proposed
 approach.
****
Separating the "Chirp" from the "Chat": Self-supervised Visual Grounding of Soun
d and Language
Mark Hamilton, Andrew Zisserman, John R. Hershey, William T. Freeman; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 13117-13127
We present DenseAV a novel dual encoder grounding architecture that learns high-
resolution semantically meaningful and audio-visual aligned features solely thro
ugh watching videos. We show that DenseAV can discover the "meaning" of words an
d the "location" of sounds without explicit localization supervision. Furthermor
e it automatically discovers and distinguishes between these two types of associ
ations without supervision. We show that DenseAV's localization abilities arise
from a new multi-head feature aggregation operator that directly compares dense
image and audio representations for contrastive learning. In contrast many other
 systems that learn "global" audio and video representations cannot localize wor
ds and sound. Finally we contribute two new datasets to improve the evaluation o
f AV representations through speech and sound prompted semantic segmentation. On
 these and other datasets we show DenseAV dramatically outperforms the prior art
 on speech and sound prompted semantic segmentation. DenseAV outperforms the cur
rent state-of-the-art ImageBind on cross-modal retrieval using fewer than half o
f the parameters. Project Page: https://aka.ms/denseav
****
Towards Generalizing to Unseen Domains with Few Labels
Chamuditha Jayanga Galappaththige, Sanoojan Baliah, Malitha Gunawardhana, Muhamm
ad Haris Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2024, pp. 23691-23700
We approach the challenge of addressing semi-supervised domain generalization (S

SDG). Specifically our aim is to obtain a model that learns domain-generalizable
 features by leveraging a limited subset of labelled data alongside a substantia
lly larger pool of unlabeled data. Existing domain generalization (DG) methods w
hich are unable to exploit unlabeled data perform poorly compared to semi-superv
ised learning (SSL) methods under SSDG setting. Nevertheless SSL methods have co
nsiderable room for performance improvement when compared to fully-supervised DG
 training. To tackle this underexplored yet highly practical problem of SSDG we
make the following core contributions. First we propose a feature-based conformi
ty technique that matches the posterior distributions from the feature space wit
h the pseudo-label from the model's output space. Second we develop a semantics
alignment loss to learn semantically-compatible representations by regularizing
the semantic structure in the feature space. Our method is plug-and-play and can
 be readily integrated with different SSL-based SSDG baselines without introduci
ng any additional parameters. Extensive experimental results across five challen
ging DG benchmarks with four strong SSL baselines suggest that our method provid
es consistent and notable gains in two different SSDG settings.
****
MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understandin
g
Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhina
v Shrivastava, Ser-Nam Lim; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2024, pp. 13504-13514
With the success of large language models (LLMs) integrating the vision model in
to LLMs to build vision-language foundation models has gained much more interest
 recently. However existing LLM-based large multimodal models (e.g. Video-LLaMA
VideoChat) can only take in a limited number of frames for short video understan
ding. In this study we mainly focus on designing an efficient and effective mode
l for long-term video understanding. Instead of trying to process more frames si
multaneously like most existing work we propose to process videos in an online m
anner and store past video information in a memory bank. This allows our model t
o reference historical video content for long-term analysis without exceeding LL
Ms' context length constraints or GPU memory limits. Our memory bank can be seam
lessly integrated into current multimodal LLMs in an off-the-shelf manner. We co
nduct extensive experiments on various video understanding tasks such as long-vi
deo understanding video question answering and video captioning and our model ca
n achieve state-of-the-art performances across multiple datasets.
****
AAMDM: Accelerated Auto-regressive Motion Diffusion Model
Tianyu Li, Calvin Qiao, Guanqiao Ren, KangKang Yin, Sehoon Ha; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 1813-1823
Interactive motion synthesis is essential in creating immersive experiences in e
ntertainment applications such as video games and virtual reality. However gener
ating animations that are both high-quality and contextually responsive remains
a challenge. Traditional techniques in the game industry can produce high-fideli
ty animations but suffer from high computational costs and poor scalability. Tra
ined neural network models alleviate the memory and speed issues yet fall short
on generating diverse motions. Diffusion models offer diverse motion synthesis w
ith low memory usage but require expensive reverse diffusion processes. This pap
er introduces the Accelerated Auto-regressive Motion Diffusion Model (AAMDM) a n
ovel motion synthesis framework designed to achieve quality diversity and effici
ency all together. AAMDM integrates Denoising Diffusion GANs as a fast Generatio
n Module and an Auto-regressive Diffusion Model as a Polishing Module. Furthermo
re AAMDM operates in a lower-dimensional embedded space rather than the full-dim
ensional pose space which reduces the training complexity as well as further imp
roves the performance. We show that AAMDM outperforms existing methods in motion
 quality diversity and runtime efficiency through comprehensive quantitative ana
lyses and visual comparisons. We also demonstrate the effectiveness of each algo
rithmic component through ablation studies.
****

Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing
Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, Jun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7817-7826
Deep Text-to-Image Synthesis (TIS) models such as Stable Diffusion have recently gained significant popularity for creative text-to-image generation. However for domain-specific scenarios tuning-free Text-guided Image Editing (TIE) is of greater importance for application developers. This approach modifies objects or object properties in images by manipulating feature components in attention layers during the generation process. Nevertheless little is known about the semantic meanings that these attention layers have learned and which parts of the attention maps contribute to the success of image editing. In this paper we conduct an in-depth probing analysis and demonstrate that cross-attention maps in Stable Diffusion often contain object attribution information which can result in editing failures. In contrast self-attention maps play a crucial role in preserving the geometric and shape details of the source image during the transformation to the target image. Our analysis offers valuable insights into understanding cross and self-attention mechanisms in diffusion models. Furthermore based on our findings we propose a simplified yet more stable and efficient tuning-free procedure that modifies only the self-attention maps of specified attention layers during the denoising process. Experimental results show that our simplified method consistently surpasses the performance of popular approaches on multiple datasets.
****
Dr2Net: Dynamic Reversible Dual-Residual Networks for Memory-Efficient Finetuning
Chen Zhao, Shuming Liu, Karttikeya Mangalam, Guocheng Qian, Fatimah Zohra, Abdulmohsen Alghannam, Jitendra Malik, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15835-15844
Large pretrained models are increasingly crucial in modern computer vision tasks. These models are typically used in downstream tasks by end-to-end finetuning which is highly memory-intensive for tasks with high-resolution data e.g. video understanding small object detection and point cloud analysis. In this paper we propose Dynamic Reversible Dual-Residual Networks or Dr2Net a novel family of network architectures that acts as a surrogate network to finetune a pretrained model with substantially reduced memory consumption. Dr2Net contains two types of residual connections one maintaining the residual structure in the pretrained models and the other making the network reversible. Due to its reversibility intermediate activations which can be reconstructed from output are cleared from memory during training. We use two coefficients on either type of residual connections respectively and introduce a dynamic training strategy that seamlessly transitions the pretrained model to a reversible network with much higher numerical precision. We evaluate Dr2Net on various pretrained models and various tasks and show that it can reach comparable performance to conventional finetuning but with significantly less memory usage.
****
PNeRV: Enhancing Spatial Consistency via Pyramidal Neural Representation for Videos
Qi Zhao, M. Salman Asif, Zhan Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19103-19112
The primary focus of Neural Representation for Videos (NeRV) is to effectively model its spatiotemporal consistency. However current NeRV systems often face a significant issue of spatial inconsistency leading to decreased perceptual quality. To address this issue we introduce the Pyramidal Neural Representation for Videos (PNeRV) which is built on a multi-scale information connection and comprises a lightweight rescaling operator Kronecker Fully-connected layer (KFc) and a Benign Selective Memory (BSM) mechanism. The KFc inspired by the tensor decomposition of the vanilla Fully-connected layer facilitates low-cost rescaling and global correlation modeling. BSM merges high-level features with granular ones adap

tively. Furthermore we provide an analysis based on the Universal Approximation Theory of the NeRV system and validate the effectiveness of the proposed PNeRV. We conducted comprehensive experiments to demonstrate that PNeRV surpasses the performance of contemporary NeRV models achieving the best results in video regression on UVG and DAVIS under various metrics (PSNR SSIM LPIPS and FVD). Compared to vanilla NeRV PNeRV achieves a +4.49 dB gain in PSNR and a 231% increase in FVD on UVG along with a +3.28 dB PSNR and 634% FVD increase on DAVIS.
****
LTGC: Long-tail Recognition via Leveraging LLMs-driven Generated Content
Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19510-19520
Long-tail recognition is challenging because it requires the model to learn good representations from tail categories and address imbalances across all categories. In this paper we propose a novel generative and fine-tuning framework LTGC to handle long-tail recognition via leveraging generated content. Firstly inspired by the rich implicit knowledge in large-scale models (e.g. large language models LLMs) LTGC leverages the power of these models to parse and reason over the original tail data to produce diverse tail-class content. We then propose several novel designs for LTGC to ensure the quality of the generated data and to efficiently fine-tune the model using both the generated and original data. The visualization demonstrates the effectiveness of the generation module in LTGC which produces accurate and diverse tail data. Additionally the experimental results demonstrate that our LTGC outperforms existing state-of-the-art methods on popular long-tailed benchmarks.
****
DiverGen: Improving Instance Segmentation by Learning Wider Data Distribution with More Diverse Generative Data
Chengxiang Fan, Muzhi Zhu, Hao Chen, Yang Liu, Weijia Wu, Huaqi Zhang, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3986-3995
Instance segmentation is data-hungry and as model capacity increases data scale becomes crucial for improving the accuracy. Most instance segmentation datasets today require costly manual annotation limiting their data scale. Models trained on such data are prone to overfitting on the training set especially for those rare categories. While recent works have delved into exploiting generative models to create synthetic datasets for data augmentation these approaches do not efficiently harness the full potential of generative models. To address these issues we introduce a more efficient strategy to construct generative datasets for data augmentation termed DiverGen. Firstly we provide an explanation of the role of generative data from the perspective of distribution discrepancy. We investigate the impact of different data on the distribution learned by the model. We argue that generative data can expand the data distribution that the model can learn thus mitigating overfitting. Additionally we find that the diversity of generative data is crucial for improving model performance and enhance it through various strategies including category diversity prompt diversity and generative model diversity. With these strategies we can scale the data to millions while maintaining the trend of model performance improvement. On the LVIS dataset DiverGen significantly outperforms the strong model X-Paste achieving +1.1 box AP and +1.1 mask AP across all categories and +1.9 box AP and +2.5 mask AP for rare categories. Our codes are available at https://github.com/aim-uofa/DiverGen.
****
Neural Refinement for Absolute Pose Regression with Feature Synthesis
Shuai Chen, Yash Bhalgat, Xinghui Li, Jia-Wang Bian, Kejie Li, Zirui Wang, Victor Adrian Prisacariu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20987-20996
Absolute Pose Regression (APR) methods use deep neural networks to directly regress camera poses from RGB images. However the predominant APR architectures only rely on 2D operations during inference resulting in limited accuracy of pose estimation due to the lack of 3D geometry constraints or priors. In this work we p

ropose a test-time refinement pipeline that leverages implicit geometric constraints using a robust feature field to enhance the ability of APR methods to use 3D information during inference. We also introduce a novel Neural Feature Synthesizer (NeFeS) model which encodes 3D geometric features during training and directly renders dense novel view features at test time to refine APR methods. To enhance the robustness of our model we introduce a feature fusion module and a progressive training strategy. Our proposed method achieves state-of-the-art single-image APR accuracy on indoor and outdoor datasets. Code will be released at https://github.com/ActiveVisionLab/NeFeS.
****
Learning Disentangled Identifiers for Action-Customized Text-to-Image Generation
Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, Donglin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7797-7806
This study focuses on a novel task in text-to-image (T2I) generation namely action customization. The objective of this task is to learn the co-existing action from limited data and generalize it to unseen humans or even animals. Experimental results show that existing subject-driven customization methods fail to learn the representative characteristics of actions and struggle in decoupling actions from context features including appearance. To overcome the preference for low-level features and the entanglement of high-level features we propose an inversion-based method Action-Disentangled Identifier (ADI) to learn action-specific identifiers from the exemplar images. ADI first expands the semantic conditioning space by introducing layer-wise identifier tokens thereby increasing the representational richness while distributing the inversion across different features. Then to block the inversion of action-agnostic features ADI extracts the gradient invariance from the constructed sample triples and masks the updates of irrelevant channels. To comprehensively evaluate the task we present an ActionBench that includes a variety of actions each accompanied by meticulously selected samples. Both quantitative and qualitative results show that our ADI outperforms existing baselines in action-customized T2I generation. Our project page is at https://adi-t2i.github.io/ADI.
****
Automatic Controllable Colorization via Imagination
Xiaoyan Cong, Yue Wu, Qifeng Chen, Chenyang Lei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2609-2619
We propose a framework for automatic colorization that allows for iterative editing and modifications. The core of our framework lies in an imagination module: by understanding the content within a grayscale image we utilize a pre-trained image generation model to generate multiple images that contain the same content. These images serve as references for coloring mimicking the process of human experts. As the synthesized images can be imperfect or different from the original grayscale image we propose a Reference Refinement Module to select the optimal reference composition. Unlike most previous end-to-end automatic colorization algorithms our framework allows for iterative and localized modifications of the colorization results because we explicitly model the coloring samples. Extensive experiments demonstrate the superiority of our framework over existing automatic colorization algorithms in editability and flexibility. Project page: https://xy-cong.github.io/imagine-colorization/.
****
Point Transformer V3: Simpler Faster Stronger
Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, Hengshuang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4840-4851
This paper is not motivated to seek innovation within the attention mechanism. Instead it focuses on overcoming the existing trade-offs between accuracy and efficiency within the context of point cloud processing leveraging the power of scale. Drawing inspiration from recent advances in 3D large-scale representation learning we recognize that model performance is more influenced by scale than by intricate design. Therefore we present Point Transformer V3 (PTv3) which prioriti

zes simplicity and efficiency over the accuracy of certain mechanisms that are minor to the overall performance after scaling such as replacing the precise neighbor search by KNN with an efficient serialized neighbor mapping of point clouds organized with specific patterns. This principle enables significant scaling expanding the receptive field from 16 to 1024 points while remaining efficient (a 3x increase in processing speed and a 10x improvement in memory efficiency compared with its predecessor PTv2). PTv3 attains state-of-the-art results on over 20 downstream tasks that span both indoor and outdoor scenarios. Further enhanced with multi-dataset joint training PTv3 pushes these results to a higher level.
****
DiffCast: A Unified Framework via Residual Diffusion for Precipitation Nowcasting
Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, Xunlai Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27758-27767
Precipitation nowcasting is an important spatio-temporal prediction task to predict the radar echoes sequences based on current observations which can serve both meteorological science and smart city applications. Due to the chaotic evolution nature of the precipitation systems it is a very challenging problem. Previous studies address the problem either from the perspectives of deterministic modeling or probabilistic modeling. However their predictions suffer from the blurry high-value echoes fading away and position inaccurate issues. The root reason of these issues is that the chaotic evolutionary precipitation systems are not appropriately modeled. Inspired by the nature of the systems we propose to decompose and model them from the perspective of global deterministic motion and local stochastic variations with residual mechanism. A unified and flexible framework that can equip any type of spatio-temporal models is proposed based on residual diffusion which effectively tackles the shortcomings of previous methods. Extensive experimental results on four publicly available radar datasets demonstrate the effectiveness and superiority of the proposed framework compared to state-of-the-art techniques. Our code is publicly available at https://github.com/DeminYu98/DiffCast.
****
Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives
Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, Michael Wray; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19383-19400
We present Ego-Exo4D a diverse large-scale multimodal multiview video dataset and benchmark challenge. Ego-Exo4D centers around simultaneously-captured egocentric and exocentric video of skilled human activities (e.g. sports music dance bike repair). 740 participants from 13 cities worldwide performed these activities

in 123 different natural scene contexts yielding long-form captures from 1 to 42 minutes each and 1286 hours of video combined. The multimodal nature of the dataset is unprecedented: the video is accompanied by multichannel audio eye gaze 3D point clouds camera poses IMU and multiple paired language descriptions---including a novel "expert commentary" done by coaches and teachers and tailored to the skilled-activity domain. To push the frontier of first-person video understanding of skilled human activity we also present a suite of benchmark tasks and their annotations including fine-grained activity understanding proficiency estimation cross-view translation and 3D hand/body pose. All resources are open sourced to fuel new research in the community.
****
Point Cloud Pre-training with Diffusion Models
Xiao Zheng, Xiaoshui Huang, Guofeng Mei, Yuenan Hou, Zhaoyang Lyu, Bo Dai, Wanli Ouyang, Yongshun Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22935-22945
Pre-training a model and then fine-tuning it on downstream tasks has demonstrated significant success in the 2D image and NLP domains. However due to the unordered and non-uniform density characteristics of point clouds it is non-trivial to explore the prior knowledge of point clouds and pre-train a point cloud backbone. In this paper we propose a novel pre-training method called Point cloud Diffusion pre-training PointDif. We consider the point cloud pre-training task as a conditional point-to-point generation problem and introduce a conditional point generator. This generator aggregates the features extracted by the backbone and employs them as the condition to guide the point-to-point recovery from the noisy point cloud thereby assisting the backbone in capturing both local and global geometric priors as well as the global point density distribution of the object. We also present a recurrent uniform sampling optimization strategy which enables the model to uniformly recover from various noise levels and learn from balanced supervision. Our PointDif achieves substantial improvement across various real-world datasets for diverse downstream tasks such as classification segmentation and detection. Specifically PointDif attains 70.0% mIoU on S3DIS Area 5 for the segmentation task and achieves an average improvement of 2.4% on ScanObjectNN for the classification task compared to TAP. Furthermore our pre-training framework can be flexibly applied to diverse point cloud backbones and bring considerable gains. Code is available at https://github.com/zhengxiaozx/PointDif
****
Mask4Align: Aligned Entity Prompting with Color Masks for Multi-Entity Localization Problems
Haoquan Zhang, Ronggang Huang, Yi Xie, Huaidong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13373-13383
In Visual Question Answering (VQA) recognizing and localizing entities pose significant challenges. Pretrained vision-and-language models have addressed this problem by providing a text description as the answer. However in visual scenes with multiple entities textual descriptions struggle to distinguish the entities from the same category effectively. Consequently the VQA dataset is limited by the limitations of text description and cannot adequately cover scenarios involving multiple entities. To address this challenge we introduce a Mask for Align (Mask4Align) method which can determine the entity's position in the given image that best matches the user-input question. This method incorporates colored masks into the image enabling the VQA model to handle discrimination and localization challenges associated with multiple entities. To process an arbitrary number of similar entities Mask4Align is designed hierarchically to discern subtle differences achieving precise localization. Since Mask4Align directly utilizes pre-trained models it does not introduce additional training overhead. Extensive experiments conducted on both the gaze target prediction task dataset and our proposed multi-entity localization dataset showcase the superiority of Mask4Align.
****
RCL: Reliable Continual Learning for Unified Failure Detection
Fei Zhu, Zhen Cheng, Xu-Yao Zhang, Cheng-Lin Liu, Zhaoxiang Zhang; Proceedings o

f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12140-12150
Deep neural networks are known to be overconfident for what they don't know in the wild which is undesirable for decision-making in high-stakes applications. Despite quantities of existing works most of them focus on detecting out-of-distribution (OOD) samples from unseen classes while ignoring large parts of relevant failure sources like misclassified samples from known classes. In particular recent studies reveal that prevalent OOD detection methods are actually harmful for misclassification detection (MisD) indicating that there seems to be a tradeoff between those two tasks. In this paper we study the critical yet under-explored problem of unified failure detection which aims to detect both misclassified and OOD examples. Concretely we identify the failure of simply integrating learning objectives of misclassification and OOD detection and show the potential of sequence learning. Inspired by this we propose a reliable continual learning paradigm whose spirit is to equip the model with MisD ability first and then improve the OOD detection ability without degrading the already adequate MisD performance. Extensive experiments demonstrate that our method achieves strong unified failure detection performance. The code is available at https://github.com/Impression2805/RCL.
****
Referring Image Editing: Object-level Image Editing via Referring Expressions
Chang Liu, Xiangtai Li, Henghui Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13128-13138
Significant advancements have been made in image editing with the recent advance of the Diffusion model. However most of the current methods primarily focus on global or subject-level modifications and often face limitations when it comes to editing specific objects when there are other objects coexisting in the scene given solely textual prompts. In response to this challenge we introduce an object-level generative task called Referring Image Editing (RIE) which enables the identification and editing of specific source objects in an image using text prompts. To tackle this task effectively we propose a tailored framework called RefDiffusion. It aims to disentangle input prompts into multiple embeddings and employs a mixed-supervised multi-stage training strategy. To facilitate further research in this domain we introduce the RefCOCO-Edit dataset comprising images editing prompts source object segmentation masks and reference edited images for training and evaluation. Our extensive experiments demonstrate the effectiveness of our approach in identifying and editing target objects while conventional general image editing and region-based image editing methods have difficulties in this challenging task.
****
CAMixerSR: Only Details Need More "Attention"
Yan Wang, Yi Liu, Shijie Zhao, Junlin Li, Li Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25837-25846
To satisfy the rapidly increasing demands on the large image (2K-8K) super-resolution (SR) prevailing methods follow two independent tracks: 1) accelerate existing networks by content-aware routing and 2) design better super-resolution networks via token mixer refining. Despite directness they encounter unavoidable defects (e.g. inflexible route or non-discriminative processing) limiting further improvements of quality-complexity trade-off. To erase the drawbacks we integrate these schemes by proposing a content-aware mixer (CAMixer) which assigns convolution for simple contexts and additional deformable window-attention for sparse textures. Specifically the CAMixer uses a learnable predictor to generate multiple bootstraps including offsets for windows warping a mask for classifying windows and convolutional attentions for endowing convolution with the dynamic property which modulates attention to include more useful textures self-adaptively and improves the representation capability of convolution. We further introduce a global classification loss to improve the accuracy of predictors. By simply stacking CAMixers we obtain CAMixerSR which achieves superior performance on large-image SR lightweight SR and omnidirectional-image SR.

****
Towards Backward-Compatible Continual Learning of Image Compression
Zhihao Duan, Ming Lu, Justin Yang, Jiangpeng He, Zhan Ma, Fengqing Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25564-25573
This paper explores the possibility of extending the capability of pre-trained neural image compressors (e.g. adapting to new data or target bitrates) without breaking backward compatibility the ability to decode bitstreams encoded by the original model. We refer to this problem as continual learning of image compression. Our initial findings show that baseline solutions such as end-to-end fine-tuning do not preserve the desired backward compatibility. To tackle this we propose a knowledge replay training strategy that effectively addresses this issue. We also design a new model architecture that enables more effective continual learning than existing baselines. Experiments are conducted for two scenarios: data-incremental learning and rate-incremental learning. The main conclusion of this paper is that neural image compressors can be fine-tuned to achieve better performance (compared to their pre-trained version) on new data and rates without compromising backward compatibility. The code is publicly available online.
****
Latent Modulated Function for Computational Optimal Continuous Image Representation
Zongyao He, Zhi Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26026-26035
The recent work Local Implicit Image Function (LIIF) and subsequent Implicit Neural Representation (INR) based works have achieved remarkable success in Arbitrary-Scale Super-Resolution (ASSR) by using MLP to decode Low-Resolution (LR) features. However these continuous image representations typically implement decoding in High-Resolution (HR) High-Dimensional (HD) space leading to a quadratic increase in computational cost and seriously hindering the practical applications of ASSR. To tackle this problem we propose a novel Latent Modulated Function (LMF) which decouples the HR-HD decoding process into shared latent decoding in LR-HD space and independent rendering in HR Low-Dimensional (LD) space thereby realizing the first computational optimal paradigm of continuous image representation. Specifically LMF utilizes an HD MLP in latent space to generate latent modulations of each LR feature vector. This enables a modulated LD MLP in render space to quickly adapt to any input feature vector and perform rendering at arbitrary resolution. Furthermore we leverage the positive correlation between modulation intensity and input image complexity to design a Controllable Multi-Scale Rendering (CMSR) algorithm offering the flexibility to adjust the decoding efficiency based on the rendering precision. Extensive experiments demonstrate that converting existing INR-based ASSR methods to LMF can reduce the computational cost by up to 99.9% accelerate inference by up to 57x and save up to 76% of parameters while maintaining competitive performance. The code is available at https://github.com/HeZongyao/LMF.
****
Unsupervised Video Domain Adaptation with Masked Pre-Training and Collaborative Self-Training
Arun Reddy, William Paul, Corban Rivera, Ketul Shah, Celso M. de Melo, Rama Chellappa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18919-18929
In this work we tackle the problem of unsupervised domain adaptation (UDA) for video action recognition. Our approach which we call UNITE uses an image teacher model to adapt a video student model to the target domain. UNITE first employs self-supervised pre-training to promote discriminative feature learning on target domain videos using a teacher-guided masked distillation objective. We then perform self-training on masked target data using the video student model and image teacher model together to generate improved pseudolabels for unlabeled target videos. Our self-training process successfully leverages the strengths of both models to achieve strong transfer performance across domains. We evaluate our approach on multiple video domain adaptation benchmarks and observe significant impr

ovements upon previously reported results.
****
UniDepth: Universal Monocular Metric Depth Estimation
Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10106-10116
Accurate monocular metric depth estimation (MMDE) is crucial to solving downstream tasks in 3D perception and modeling. However the remarkable accuracy of recent MMDE methods is confined to their training domains. These methods fail to generalize to unseen domains even in the presence of moderate domain gaps which hinders their practical applicability. We propose a new model UniDepth capable of reconstructing metric 3D scenes from solely single images across domains. Departing from the existing MMDE methods UniDepth directly predicts metric 3D points from the input image at inference time without any additional information striving for a universal and flexible MMDE solution. In particular UniDepth implements a self-promptable camera module predicting dense camera representation to condition depth features. Our model exploits a pseudo-spherical output representation which disentangles camera and depth representations. In addition we propose a geometric invariance loss that promotes the invariance of camera-prompted depth features. Thorough evaluations on ten datasets in a zero-shot regime consistently demonstrate the superior performance of UniDepth even when compared with methods directly trained on the testing domains. Code and models are available at: github.com/lpiccinelli-eth/unidepth
****
EMOPortraits: Emotion-enhanced Multimodal One-shot Head Avatars
Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, Maja Pantic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8498-8507
Head avatars animated by visual signals have gained popularity particularly in cross-driving synthesis where the driver differs from the animated character a challenging but highly practical approach. The recently presented MegaPortraits model has demonstrated state-of-the-art results in this domain. We conduct a deep examination and evaluation of this model with a particular focus on its latent space for facial expression descriptors and uncover several limitations with its ability to express intense face motions. Head avatars animated by visual signals have gained popularity particularly in cross-driving synthesis where the driver differs from the animated character a challenging but highly practical approach. The recently presented MegaPortraits model has demonstrated state-of-the-art results in this domain. We conduct a deep examination and evaluation of this model with a particular focus on its latent space for facial expression descriptors and uncover several limitations with its ability to express intense face motions. To address these limitations we propose substantial changes in both training pipeline and model architecture to introduce our EMOPortraits model where we: Enhance the model's capability to faithfully support intense asymmetric face expressions setting a new state-of-the-art result in the emotion transfer task surpassing previous methods in both metrics and quality. Incorporate speech-driven mode to our model achieving top-tier performance in audio-driven facial animation making it possible to drive source identity through diverse modalities including visual signal audio or a blend of both.Furthermore we propose a novel multi-view video dataset featuring a wide range of intense and asymmetric facial expressions filling the gap with absence of such data in existing datasets.
****
NeuRAD: Neural Rendering for Autonomous Driving
Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, Christoffer Petersson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14895-14904
Neural radiance fields (NeRFs) have gained popularity in the autonomous driving (AD) community. Recent methods show NeRFs' potential for closed-loop simulation enabling testing of AD systems and as an advanced training data augmentation technique. However existing methods often require long training times dense semanti

c supervision or lack generalizability. This in turn hinders the application of NeRFs for AD at scale. In this paper we propose \modelname a robust novel view synthesis method tailored to dynamic AD data. Our method features simple network design extensive sensor modeling for both camera and lidar -- including rolling shutter beam divergence and ray dropping -- and is applicable to multiple datasets out of the box. We verify its performance on five popular AD datasets achieving state-of-the-art performance across the board. To encourage further development we openly release the NeuRAD source code at https://github.com/georghess/NeuRAD.
****
VideoCutLER: Surprisingly Simple Unsupervised Video Instance Segmentation
Xudong Wang, Ishan Misra, Ziyun Zeng, Rohit Girdhar, Trevor Darrell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22755-22764
Existing approaches to unsupervised video instance segmentation typically rely on motion estimates and experience difficulties tracking small or divergent motions. We present VideoCutLER a simple method for unsupervised multi-instance video segmentation without using motion-based learning signals like optical flow or training on natural videos. Our key insight is that using high-quality pseudo masks and a simple video synthesis method for model training is surprisingly sufficient to enable the resulting video model to effectively segment and track multiple instances across video frames. We show the first competitive unsupervised learning results on the challenging YouTubeVIS-2019 benchmark achieving 50.7% AP50 surpassing the previous state-of-the-art by a large margin. VideoCutLER can also serve as a strong pretrained model for supervised video instance segmentation tasks exceeding DINO by 15.9% on YouTubeVIS-2019 in terms of AP.
****
Bootstrapping Chest CT Image Understanding by Distilling Knowledge from X-ray Expert Models
Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony C. W. Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, Ling Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11238-11247
Radiologists highly desire fully automated versatile AI for medical imaging interpretation. However the lack of extensively annotated large-scale multi-disease datasets has hindered the achievement of this goal. In this paper we explore the feasibility of leveraging language as a naturally high-quality supervision for chest CT imaging. In light of the limited availability of image-report pairs we bootstrap the understanding of 3D chest CT images by distilling chest-related diagnostic knowledge from an extensively pre-trained 2D X-ray expert model. Specifically we propose a language-guided retrieval method to match each 3D CT image with its semantically closest 2D X-ray image and perform pair-wise and semantic relation knowledge distillation. Subsequently we use contrastive learning to align images and reports within the same patient while distinguishing them from the other patients. However the challenge arises when patients have similar semantic diagnoses such as healthy patients potentially confusing if treated as negatives. We introduce a robust contrastive learning that identifies and corrects these false negatives. We train our model with over 12K pairs of chest CT images and radiology reports. Extensive experiments across multiple scenarios including zero-shot learning report generation and fine-tuning processes demonstrate the model's feasibility in interpreting chest CT images.
****
Magic Tokens: Select Diverse Tokens for Multi-modal Object Re-Identification
Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17117-17126
Single-modal object re-identification (ReID) faces great challenges in maintaining robustness within complex visual scenarios. In contrast multi-modal object ReID utilizes complementary information from diverse modalities showing great potentials for practical applications. However previous methods may be easily affected by irrelevant backgrounds and usually ignore the modality gaps. To address ab

ove issues we propose a novel learning framework named EDITOR to select diverse tokens from vision Transformers for multi-modal object ReID. We begin with a shared vision Transformer to extract tokenized features from different input modalities. Then we introduce a Spatial-Frequency Token Selection (SFTS) module to adaptively select object-centric tokens with both spatial and frequency information. Afterwards we employ a Hierarchical Masked Aggregation (HMA) module to facilitate feature interactions within and across modalities. Finally to further reduce the effect of backgrounds we propose a Background Consistency Constraint (BCC) and an Object-Centric Feature Refinement (OCFR). They are formulated as two new loss functions which improve the feature discrimination with background suppression. As a result our framework can generate more discriminative features for multi-modal object ReID. Extensive experiments on three multi-modal ReID benchmarks verify the effectiveness of our methods. The code is available at https://github.com/924973292/EDITOR.
****
Open3DIS: Open-Vocabulary 3D Instance Segmentation with 2D Mask Guidance
Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, Khoi Nguyen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4018-4028
We introduce Open3DIS a novel solution designed to tackle the problem of Open-Vocabulary Instance Segmentation within 3D scenes. Objects within 3D environments exhibit diverse shapes scales and colors making precise instance-level identification a challenging task. Recent advancements in Open-Vocabulary scene understanding have made significant strides in this area by employing class-agnostic 3D instance proposal networks for object localization and learning queryable features for each 3D mask. While these methods produce high-quality instance proposals they struggle with identifying small-scale and geometrically ambiguous objects. The key idea of our method is a new module that aggregates 2D instance masks across frames and maps them to geometrically coherent point cloud regions as high-quality object proposals addressing the above limitations. These are then combined with 3D class-agnostic instance proposals to include a wide range of objects in the real world. To validate our approach we conducted experiments on three prominent datasets including ScanNet200 S3DIS and Replica demonstrating significant performance gains in segmenting objects with diverse categories over the state-of-the-art approaches.
****
SignGraph: A Sign Sequence is Worth Graphs of Nodes
Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Hongkai Wen, Lei Xie, Sanglu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13470-13479
Despite the recent success of sign language research the widely adopted CNN-based backbones are mainly migrated from other computer vision tasks in which the contours and texture of objects are crucial for identifying objects. They usually treat sign frames as grids and may fail to capture effective cross-region features. In fact sign language tasks need to focus on the correlation of different regions in one frame and the interaction of different regions among adjacent frames for identifying a sign sequence. In this paper we propose to represent a sign sequence as graphs and introduce a simple yet effective graph-based sign language processing architecture named SignGraph to extract cross-region features at the graph level. SignGraph consists of two basic modules: Local Sign Graph (LSG) module for learning the correlation of intra-frame cross-region features in one frame and Temporal Sign Graph (TSG) module for tracking the interaction of inter-frame cross-region features among adjacent frames. With LSG and TSG we build our model in a multiscale manner to ensure that the representation of nodes can capture cross-region features at different granularities. Extensive experiments on current public sign language datasets demonstrate the superiority of our SignGraph model. Our model achieves very competitive performances with the SOTA model while not using any extra cues. Code and models are available at: https://github.com/gswycf/SignGraph.
****

ControlRoom3D: Room Generation using Semantic Proxy Rooms
Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, Peizhao Zhang, Bastian Leibe, Peter Vajda, Ji Hou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6201-6210
Manually creating 3D environments for AR/VR applications is a complex process requiring expert knowledge in 3D modeling software. Pioneering works facilitate this process by generating room meshes conditioned on textual style descriptions. Yet many of these automatically generated 3D meshes do not adhere to typical room layouts compromising their plausibility e.g. by placing several beds in one bedroom. To address these challenges we present ControlRoom3D a novel method to generate high-quality room meshes. Central to our approach is a user-defined 3D semantic proxy room that outlines a rough room layout based on semantic bounding boxes and a textual description of the overall room style. Our key insight is that when rendered to 2D this 3D representation provides valuable geometric and semantic information to control powerful 2D models to generate 3D consistent textures and geometry that aligns well with the proxy room. Backed up by an extensive study including quantitative metrics and qualitative user evaluations our method generates diverse and globally plausible 3D room meshes thus empowering users to design 3D rooms effortlessly without specialized knowledge.
****
DeconfuseTrack: Dealing with Confusion for Multi-Object Tracking
Cheng Huang, Shoudong Han, Mengyu He, Wenbo Zheng, Yuhao Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19290-19299
Accurate data association is crucial in reducing confusion such as ID switches and assignment errors in multi-object tracking (MOT). However existing advanced methods often overlook the diversity among trajectories and the ambiguity and conflicts present in motion and appearance cues leading to confusion among detections trajectories and associations when performing simple global data association. To address this issue we propose a simple versatile and highly interpretable data association approach called Decomposed Data Association (DDA). DDA decomposes the traditional association problem into multiple sub-problems using a series of non-learning-based modules and selectively addresses the confusion in each sub-problem by incorporating targeted exploitation of new cues. Additionally we introduce Occlusion-aware Non-Maximum Suppression (ONMS) to retain more occluded detections thereby increasing opportunities for association with trajectories and indirectly reducing the confusion caused by missed detections. Finally based on DDA and ONMS we design a powerful multi-object tracker named DeconfuseTrack specifically focused on resolving confusion in MOT. Extensive experiments conducted on the MOT17 and MOT20 datasets demonstrate that our proposed DDA and ONMS significantly enhance the performance of several popular trackers. Moreover DeconfuseTrack achieves state-of-the-art performance on the MOT17 and MOT20 test sets significantly outperforms the baseline tracker ByteTrack in metrics such as HOTA IDF1 AssA. This validates that our tracking design effectively reduces confusion caused by simple global association.
****
PAPR in Motion: Seamless Point-level 3D Scene Interpolation
Shichong Peng, Yanshu Zhang, Ke Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21007-21016
We propose the problem of point-level 3D scene interpolation which aims to simultaneously reconstruct a 3D scene in two states from multiple views synthesize smooth point-level interpolations between them and render the scene from novel viewpoints all without any supervision between the states. The primary challenge is on achieving a smooth transition between states that may involve significant and non-rigid changes. To address these challenges we introduce "PAPR in Motion" a novel approach that builds upon the recent Proximity Attention Point Rendering (PAPR) technique which can deform a point cloud to match a significantly different shape and render a visually coherent scene even after non-rigid deformations. Our approach is specifically designed to maintain the temporal consistency of t

he geometric structure by introducing various regularization techniques for PAPR. The result is a method that can effectively bridge large scene changes and produce visually coherent and temporally smooth interpolations in both geometry and appearance. Evaluation across diverse motion types demonstrates that "PAPR in Motion" outperforms the leading neural renderer for dynamic scenes. For more results and code please visit our project website at https://niopeng.github.io/PAPR-in-Motion/.
****
Causal Mode Multiplexer: A Novel Framework for Unbiased Multispectral Pedestrian Detection
Taeheon Kim, Sebin Shin, Youngjoon Yu, Hak Gu Kim, Yong Man Ro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26784-26793
RGBT multispectral pedestrian detection has emerged as a promising solution for safety-critical applications that require day/night operations. However the modality bias problem remains unsolved as multispectral pedestrian detectors learn the statistical bias in datasets. Specifically datasets in multispectral pedestrian detection mainly distribute between ROTO (day) and RXTO (night) data; the majority of the pedestrian labels statistically co-occur with their thermal features. As a result multispectral pedestrian detectors show poor generalization ability on examples beyond this statistical correlation such as ROTX data. To address this problem we propose a novel Causal Mode Multiplexer (CMM) framework that effectively learns the causalities between multispectral inputs and predictions. Moreover we construct a new dataset (ROTX-MP) to evaluate modality bias in multispectral pedestrian detection. ROTX-MP mainly includes ROTX examples not presented in previous datasets. Extensive experiments demonstrate that our proposed CMM framework generalizes well on existing datasets (KAIST CVC-14 FLIR) and the new ROTX-MP. Our code and dataset are available open-source.
****
HIMap: HybrId Representation Learning for End-to-end Vectorized HD Map Construction
Yi Zhou, Hui Zhang, Jiaqian Yu, Yifan Yang, Sangil Jung, Seung-In Park, ByungIn Yoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15396-15406
Vectorized High-Definition (HD) map construction requires predictions of the category and point coordinates of map elements (e.g. road boundary lane divider pedestrian crossing etc.). State-of-the-art methods are mainly based on point-level representation learning for regressing accurate point coordinates. However this pipeline has limitations in obtaining element-level information and handling element-level failures e.g. erroneous element shape or entanglement between elements. To tackle the above issues we propose a simple yet effective HybrId framework named HIMap to sufficiently learn and interact both point-level and element-level information. Concretely we introduce a hybrid representation called HIQuery to represent all map elements and propose a point-element interactor to interactively extract and encode the hybrid information of elements e.g. point position and element shape into the HIQuery. Additionally we present a point-element consistency constraint to enhance the consistency between the point-level and element-level information. Finally the output point-element integrated HIQuery can be directly converted into map elements' class point coordinates and mask. We conduct extensive experiments and consistently outperform previous methods on both nuScenes and Argoverse2 datasets. Notably our method achieves 77.8 mAP on the nuScenes dataset remarkably superior to previous SOTAs by 8.3 mAP at least.
****
LTA-PCS: Learnable Task-Agnostic Point Cloud Sampling
Jiaheng Liu, Jianhao Li, Kaisiyuan Wang, Hongcheng Guo, Jian Yang, Junran Peng, Ke Xu, Xianglong Liu, Jinyang Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28035-28045
Recently many approaches directly operate on point clouds for different tasks. These approaches become more computation and storage demanding when point cloud size is large. To reduce the required computation and storage one possible soluti

on is to sample the point cloud. In this paper we propose the first Learnable Ta sk-Agnostic Point Cloud Sampling (LTA-PCS) framework. Existing task-agnostic poi nt cloud sampling strategy (e.g. FPS) does not consider semantic information of point clouds causing degraded performance on downstream tasks. While learning-ba sed point cloud sampling methods consider semantic information they are task-spe cific and require task-oriented ground-truth annotations. So they cannot general ize well on different downstream tasks. Our LTA-PCS achieves task-agnostic point cloud sampling without requiring task-oriented labels in which both the geometr ic and semantic information of points is considered in sampling. Extensive exper iments on multiple downstream tasks demonstrate the effectiveness of our LTA-PCS .
****
Non-Rigid Structure-from-Motion: Temporally-Smooth Procrustean Alignment and Spa tially-Variant Deformation Modeling
Jiawei Shi, Hui Deng, Yuchao Dai; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2024, pp. 21446-21455
Even though Non-rigid Structure-from-Motion (NRSfM) has been extensively studied and great progress has been made there are still key challenges that hinder the ir broad real-world applications: 1) the inherent motion/rotation ambiguity requ ires either explicit camera motion recovery with extra constraint or complex Pro crustean Alignment; 2) existing low-rank modeling of the global shape can over-p enalize drastic deformations in the 3D shape sequence. This paper proposes to re solve the above issues from a spatial-temporal modeling perspective. First we pr opose a novel Temporally-smooth Procrustean Alignment module that estimates 3D d eforming shapes and adjusts the camera motion by aligning the 3D shape sequence consecutively. Our new alignment module remedies the requirement of complex refe rence 3D shape during alignment which is more conductive to non-isotropic deform ation modeling. Second we propose a spatial-weighted approach to enforce the low -rank constraint adaptively at different locations to accommodate drastic spatia lly-variant deformation reconstruction better. Our modeling outperform existing low-rank based methods and extensive experiments across different datasets valid ate the effectiveness of our method.
****
ShapeMatcher: Self-Supervised Joint Shape Canonicalization Segmentation Retrieva l and Deformation
Yan Di, Chenyangguang Zhang, Chaowei Wang, Ruida Zhang, Guangyao Zhai, Yanyan Li , Bowen Fu, Xiangyang Ji, Shan Gao; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2024, pp. 21017-21028
In this paper we present ShapeMatcher a unified self-supervised learning framewo rk for joint shape canonicalization segmentation retrieval and deformation. Give n a partially-observed object in an arbitrary pose we first canonicalize the obj ect by extracting point-wise affine invariant features disentangling inherent st ructure of the object with its pose and size. These learned features are then le veraged to predict semantically consistent part segmentation and corresponding p art centers. Next our lightweight retrieval module aggregates the features withi n each part as its retrieval token and compare all the tokens with source shapes from a pre-established database to identify the most geometrically similar shap e. Finally we deform the retrieved shape in the deformation module to tightly fi t the input object by harnessing part center guided neural cage deformation. The key insight of ShapeMaker is the simultaneous training of the four highly-assoc iated processes: canonicalization segmentation retrieval and deformation leverag ing cross-task consistency losses for mutual supervision. Extensive experiments on synthetic datasets PartNet ComplementMe and real-world dataset Scan2CAD demon strate that ShapeMatcher surpasses competitors by a large margin. Code is releas ed at https://github.com/Det1999/ShapeMaker.
****
UniPTS: A Unified Framework for Proficient Post-Training Sparsity
Jingjing Xie, Yuxin Zhang, Mingbao Lin, Liujuan Cao, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024 , pp. 5746-5755

Post-training Sparsity (PTS) is a recently emerged avenue that chases efficient network sparsity with limited data in need. Existing PTS methods however undergo significant performance degradation compared with traditional methods that retrain the sparse networks via the whole dataset especially at high sparsity ratios. In this paper we attempt to reconcile this disparity by transposing three cardinal factors that profoundly alter the performance of conventional sparsity into the context of PTS. Our endeavors particularly comprise (1) A base-decayed sparsity objective that promotes efficient knowledge transferring from dense network to the sparse counterpart. (2) A reducing-regrowing search algorithm designed to ascertain the optimal sparsity distribution while circumventing overfitting to the small calibration set in PTS. (3) The employment of dynamic sparse training predicated on the preceding aspects aimed at comprehensively optimizing the sparsity structure while ensuring training stability. Our proposed framework termed UniPTS is validated to be much superior to existing PTS methods across extensive benchmarks. As an illustration it amplifies the performance of POT a recently proposed recipe from 3.9% to 68.6% when pruning ResNet-50 at 90% sparsity ratio on ImageNet.
****
HumanNorm: Learning Normal Diffusion Model for High-quality and Realistic 3D Human Generation
Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, Qing Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4568-4577
Recent text-to-3D methods employing diffusion models have made significant advancements in 3D human generation. However these approaches face challenges due to the limitations of text-to-image diffusion models which lack an understanding of 3D structures. Consequently these methods struggle to achieve high-quality human generation resulting in smooth geometry and cartoon-like appearances. In this paper we propose HumanNorm a novel approach for high-quality and realistic 3D human generation. The main idea is to enhance the model's 2D perception of 3D geometry by learning a normal-adapted diffusion model and a normal-aligned diffusion model. The normal-adapted diffusion model can generate high-fidelity normal maps corresponding to user prompts with view-dependent and body-aware text. The normal-aligned diffusion model learns to generate color images aligned with the normal maps thereby transforming physical geometry details into realistic appearance. Leveraging the proposed normal diffusion model we devise a progressive geometry generation strategy and a multi-step Score Distillation Sampling (SDS) loss to enhance the performance of 3D human generation. Comprehensive experiments substantiate HumanNorm's ability to generate 3D humans with intricate geometry and realistic appearances. HumanNorm outperforms existing text-to-3D methods in both geometry and texture quality. The project page of HumanNorm is https://humannorm.github.io/.
****
Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization
Guopeng Li, Ming Qian, Gui-Song Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16719-16729
This paper investigates the effective utilization of unlabeled data for large-area cross-view geo-localization (CVGL) encompassing both unsupervised and semi-supervised settings. Common approaches to CVGL rely on ground-satellite image pairs and employ label-driven supervised training. However the cost of collecting precise cross-view image pairs hinders the deployment of CVGL in real-life scenarios. Without the pairs CVGL will be more challenging to handle the significant imaging and spatial gaps between ground and satellite images. To this end we propose an unsupervised framework including a cross-view projection to guide the model for retrieving initial pseudo-labels and a fast re-ranking mechanism to refine the pseudo-labels by leveraging the fact that "the perfectly paired ground-satellite image is located in a unique and identical scene". The framework exhibits competitive performance compared with supervised works on three open-source benchmarks. Our code and models will be released on https://github.com/liguopeng0923/UCVGL.

****
Global Latent Neural Rendering
Thomas Tanay, Matteo Maggioni; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2024, pp. 19723-19733
A recent trend among generalizable novel view synthesis methods is to learn a re ndering operator acting over single camera rays. This approach is promising beca use it removes the need for explicit volumetric rendering but it effectively tre ats target images as collections of independent pixels. Here we propose to learn a global rendering operator acting over all camera rays jointly. We show that t he right representation to enable such rendering is a 5-dimensional plane sweep volume consisting of the projection of the input images on a set of planes facin g the target camera. Based on this understanding we introduce our Convolutional Global Latent Renderer (ConvGLR) an efficient convolutional architecture that pe rforms the rendering operation globally in a low-resolution latent space. Experi ments on various datasets under sparse and generalizable setups show that our ap proach consistently outperforms existing methods by significant margins.
****
PanoOcc: Unified Occupancy Representation for Camera-based 3D Panoptic Segmentat ion
Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, Zhaoxiang Zhang; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p p. 17158-17168
Comprehensive modeling of the surrounding 3D world is crucial for the success of autonomous driving. However existing perception tasks like object detection roa d structure segmentation depth & elevation estimation and open-set object locali zation each only focus on a small facet of the holistic 3D scene understanding t ask. This divide-and-conquer strategy simplifies the algorithm development proce ss but comes at the cost of losing an end-to-end unified solution to the problem . In this work we address this limitation by studying camera-based 3D panoptic s egmentation aiming to achieve a unified occupancy representation for camera-only 3D scene understanding. To achieve this we introduce a novel method called Pano Occ which utilizes voxel queries to aggregate spatiotemporal information from mu lti-frame and multi-view images in a coarse-to-fine scheme integrating feature l earning and scene representation into a unified occupancy representation. We hav e conducted extensive ablation studies to validate the effectiveness and efficie ncy of the proposed method. Our approach achieves new state-of-the-art results f or camera-based semantic segmentation and panoptic segmentation on the nuScenes dataset. Furthermore our method can be easily extended to dense occupancy predic tion and has demonstrated promising performance on the Occ3D benchmark. The code will be made available at https://github.com/Robertwyq/PanoOcc.
****
Sparse Views Near Light: A Practical Paradigm for Uncalibrated Point-light Photo metric Stereo
Mohammed Brahimi, Bjoern Haefner, Zhenzhang Ye, Bastian Goldluecke, Daniel Creme rs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2024, pp. 11862-11872
Neural approaches have shown a significant progress on camera-based reconstructi on. But they require either a fairly dense sampling of the viewing sphere or pre -training on an existing dataset thereby limiting their generalizability. In con trast photometric stereo (PS) approaches have shown great potential for achievin g high-quality reconstruction under sparse viewpoints. Yet they are impractical because they typically require tedious laboratory conditions are restricted to d ark rooms and often multi-staged making them subject to accumulated errors. To a ddress these shortcomings we propose an end-to-end uncalibrated multi-view PS fr amework for reconstructing high-resolution shapes acquired from sparse viewpoint s in a real-world environment. We relax the dark room assumption and allow a com bination of static ambient lighting and dynamic near LED lighting thereby enabli ng easy data capture outside the lab. Experimental validation confirms that it o utperforms existing baseline approaches in the regime of sparse viewpoints by a large margin. This allows to bring high accuracy 3D reconstruction from the dark

room to the real world while maintaining a reasonable data capture complexity.
****
Meta-Point Learning and Refining for Category-Agnostic Pose Estimation
Junjie Chen, Jiebin Yan, Yuming Fang, Li Niu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23534-23543
Category-agnostic pose estimation (CAPE) aims to predict keypoints for arbitrary classes given a few support images annotated with keypoints. Existing methods only rely on the features extracted at support keypoints to predict or refine the keypoints on query image but a few support feature vectors are local and inadequate for CAPE. Considering that human can quickly perceive potential keypoints of arbitrary objects we propose a novel framework for CAPE based on such potential keypoints (named as meta-points). Specifically we maintain learnable embeddings to capture inherent information of various keypoints which interact with image feature maps to produce meta-points without any support. The produced meta-points could serve as meaningful potential keypoints for CAPE. Due to the inevitable gap between inherency and annotation we finally utilize the identities and details offered by support keypoints to assign and refine meta-points to desired keypoints in query image. In addition we propose a progressive deformable point decoder and a slacked regression loss for better prediction and supervision. Our novel framework not only reveals the inherency of keypoints but also outperforms existing methods of CAPE. Comprehensive experiments and in-depth studies on large-scale MP-100 dataset demonstrate the effectiveness of our framework.
****
Cross-view and Cross-pose Completion for 3D Human Understanding
Matthieu Armando, Salma Galaaoui, Fabien Baradel, Thomas Lucas, Vincent Leroy, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1512-1523
Human perception and understanding is a major domain of computer vision which like many other vision subdomains recently stands to gain from the use of large models pre-trained on large datasets. We hypothesize that the most common pre-training strategy of relying on general purpose object-centric image datasets such as ImageNet is limited by an important domain shift. On the other hand collecting domain-specific ground truth such as 2D or 3D labels does not scale well. Therefore we propose a pre-training approach based on self-supervised learning that works on human-centric data using only images. Our method uses pairs of images of humans: the first is partially masked and the model is trained to reconstruct the masked parts given the visible ones and a second image. It relies on both stereoscopic (cross-view) pairs and temporal (cross-pose) pairs taken from videos in order to learn priors about 3D as well as human motion. We pre-train a model for body-centric tasks and one for hand-centric tasks. With a generic transformer architecture these models outperform existing self-supervised pre-training methods on a wide set of human-centric downstream tasks and obtain state-of-the-art performance for instance when fine-tuning for model-based and model-free human mesh recovery.
****
Batch Normalization Alleviates the Spectral Bias in Coordinate Networks
Zhicheng Cai, Hao Zhu, Qiu Shen, Xinran Wang, Xun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25160-25171
Representing signals using coordinate networks dominates the area of inverse problems recently and is widely applied in various scientific computing tasks. Still there exists an issue of spectral bias in coordinate networks limiting the capacity to learn high-frequency components. This problem is caused by the pathological distribution of the neural tangent kernel's (NTK's) eigenvalues of coordinate networks. We find that this pathological distribution could be improved using the classical batch normalization (BN) which is a common deep learning technique but rarely used in coordinate networks. BN greatly reduces the maximum and variance of NTK's eigenvalues while slightly modifies the mean value considering the max eigenvalue is much larger than the most this variance change results in a

shift of eigenvalues' distribution from a lower one to a higher one therefore th
e spectral bias could be alleviated (see Fig. 1). This observation is substantia
ted by the significant improvements of applying BN-based coordinate networks to
various tasks including the image compression computed tomography reconstruction
 shape representation magnetic resonance imaging and novel view synthesis.
****
Efficient Scene Recovery Using Luminous Flux Prior
Zhongyu Li, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2024, pp. 2743-2752
Scene recovery the restoration of images degraded by adverse weather conditions
presents significant challenges for existing methods. Physical models constraine
d by their inherent assumptions often fail when these assumptions are not met; D
eep learning models are powerful they are limited by the diversity of their trai
ning datasets leading to poor generalization and high computational demands. To
address these limitations we propose the Luminous Flux Prior (LFP) to recover de
graded images under diverse adverse weather without learning. Luminous flux a ph
ysical measure that reflects image brightness has a rate of change that demonstr
ates a significant correlation with transmission. Consequently we leverage this
rate of change in luminous flux as prior knowledge to estimate transmission whic
h in turn assists in image recovery. This approach reduces dependency on physica
l parameters and enhances adaptability to various weather. Experimental validati
on under diverse conditions such as sandstorms underwater environments and haze
attests to the robustness of LFP in restoring clear images. With a time complexi
ty of \mathcal O (N\log N) LFP enables real-time recovery making it a suitable f
or devices with limited computational resources.
****
LQMFormer: Language-aware Query Mask Transformer for Referring Image Segmentatio
n
Nisarg A. Shah, Vibashan VS, Vishal M. Patel; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12903-12913
Referring Image Segmentation (RIS) aims to segment objects from an image based o
n a language description. Recent advancements have introduced transformer-based
methods that leverage cross-modal dependencies significantly enhancing performan
ce in referring segmentation tasks. These methods are designed such that each qu
ery predicts different masks. However RIS inherently requires a single-mask pred
iction leading to a phenomenon known as Query Collapse where all queries yield t
he same mask prediction. This reduces the generalization capability of the RIS m
odel for complex or novel scenarios. To address this issue we propose a Multi-mo
dal Query Feature Fusion technique characterized by two innovative designs: (1)
Gaussian enhanced Multi-Modal Fusion a novel visual grounding mechanism that enh
ances overall representation by extracting rich local visual information and glo
bal visual-linguistic relationships and (2) A Dynamic Query Module that produces
 a diverse set of queries through a scoring network where the network selectivel
y focuses on queries for objects referred to in the language description. Moreov
er we show that including an auxiliary loss to increase the distance between mas
k representations of different queries further enhances performance and mitigate
s query collapse. Extensive experiments conducted on four benchmark datasets val
idate the effectiveness of our framework.
****
Customize your NeRF: Adaptive Source Driven 3D Scene Editing via Local-Global It
erative Training
Runze He, Shaofei Huang, Xuecheng Nie, Tianrui Hui, Luoqi Liu, Jiao Dai, Jizhong
 Han, Guanbin Li, Si Liu; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2024, pp. 6966-6975
In this paper we target the adaptive source driven 3D scene editing task by prop
osing a CustomNeRF model that unifies a text description or a reference image as
 the editing prompt. However obtaining desired editing results conformed with th
e editing prompt is nontrivial since there exist two significant challenges incl
uding accurate editing of only foreground regions and multi-view consistency giv
en a single-view reference image. To tackle the first challenge we propose a Loc

al-Global Iterative Editing (LGIE) training scheme that alternates between foreg round region editing and full-image editing aimed at foreground-only manipulatio n while preserving the background. For the second challenge we also design a cla ss-guided regularization that exploits class priors within the generation model to alleviate the inconsistency problem among different views in image-driven edi ting. Extensive experiments show that our CustomNeRF produces precise editing re sults under various real scenes for both text- and image-driven settings. The co de is available at: https://github. com/hrz2000/CustomNeRF.
****
SplaTAM: Splat Track & Map 3D Gaussians for Dense RGB-D SLAM
Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebast ian Scherer, Deva Ramanan, Jonathon Luiten; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21357-21366
Dense simultaneous localization and mapping (SLAM) is crucial for robotics and a ugmented reality applications. However current methods are often hampered by the non-volumetric or implicit way they represent a scene. This work introduces Spl aTAM an approach that for the first time leverages explicit volumetric represent ations i.e. 3D Gaussians to enable high-fidelity reconstruction from a single un posed RGB-D camera surpassing the capabilities of existing methods. SplaTAM empl oys a simple online tracking and mapping system tailored to the underlying Gauss ian representation. It utilizes a silhouette mask to elegantly capture the prese nce of scene density. This combination enables several benefits over prior repre sentations including fast rendering and dense optimization quickly determining i f areas have been previously mapped and structured map expansion by adding more Gaussians. Extensive experiments show that SplaTAM achieves up to 2x superior pe rformance in camera pose estimation map construction and novel-view synthesis ov er existing methods paving the way for more immersive high-fidelity SLAM applica tions.
****
Instance-based Max-margin for Practical Few-shot Recognition
Minghao Fu, Ke Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2024, pp. 28674-28683
In order to mimic the human few-shot learning (FSL) ability better and to make F SL closer to real-world applications this paper proposes a practical FSL (pFSL) setting. pFSL is based on unsupervised pre-trained models (analogous to human pr ior knowledge) and recognizes many novel classes simultaneously. Compared to tra ditional FSL pFSL is simpler in its formulation easier to evaluate more challeng ing and more practical. To cope with the rarity of training examples this paper proposes IbM2 an instance-based max-margin method not only for the new pFSL sett ing but also works well in traditional FSL scenarios. Based on the Gaussian Annu lus Theorem IbM2 converts random noise applied to the instances into a mechanism to achieve maximum margin in the many-way pFSL (or traditional FSL) recognition task. Experiments with various self-supervised pre-training methods and diverse many- or few-way FSL tasks show that IbM2 almost always leads to improvements c ompared to its respective baseline methods and in most cases the improvements ar e significant. With both the new pFSL setting and novel IbM2 method this paper s hows that practical few-shot learning is both viable and promising.
****
Spherical Mask: Coarse-to-Fine 3D Point Cloud Instance Segmentation with Spheric al Representation
Sangyun Shin, Kaichen Zhou, Madhu Vankadari, Andrew Markham, Niki Trigoni; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2024, pp. 4060-4069
Coarse-to-fine 3D instance segmentation methods show weak performances compared to recent Grouping-based Kernel-based and Transformer-based methods. We argue th at this is due to two limitations: 1) Instance size overestimation by axis-align ed bounding box(AABB) 2) False negative error accumulation from inaccurate box t o the refinement phase. In this work we introduce Spherical Mask a novel coarse- to-fine approach based on spherical representation overcoming those two limitati ons with several benefits. Specifically our coarse detection estimates each inst

ance with a 3D polygon using a center and radial distance predictions which avoids excessive size estimation of AABB. To cut the error propagation in the existing coarse-to-fine approaches we virtually migrate points based on the polygon allowing all foreground points including false negatives to be refined. During inference the proposal and point migration modules run in parallel and are assembled to form binary masks of instances. We also introduce two margin-based losses for the point migration to enforce corrections for the false positives/negatives and cohesion of foreground points significantly improving the performance. Experimental results from three datasets such as ScanNetV2 S3DIS and STPLS3D show that our proposed method outperforms existing works demonstrating the effectiveness of the new instance representation with spherical coordinates. The code is available at: https://github.com/yunshin/SphericalMask
****
Omni-Q: Omni-Directional Scene Understanding for Unsupervised Visual Grounding
Sai Wang, Yutian Lin, Yu Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14261-14270
Unsupervised visual grounding methods alleviate the issue of expensive manual annotation of image-query pairs by generating pseudo-queries. However existing methods are prone to confusing the spatial relationships between objects and rely on designing complex prompt modules to generate query texts which severely impedes the ability to generate accurate and comprehensive queries due to ambiguous spatial relationships and manually-defined fixed templates. To tackle these challenges we propose a omni-directional language query generation approach for unsupervised visual grounding named Omni-Q. Specifically we develop a 3D spatial relation module to extend the 2D spatial representation to 3D thereby utilizing 3D location information to accurately determine the spatial position among objects. Besides we introduce a spatial graph module leveraging the power of graph structures to establish accurate and diverse object relationships and thus enhancing the flexibility of query generation. Extensive experiments on five public benchmark datasets demonstrate that our method significantly outperforms existing state-of-the-art unsupervised methods by up to 16.17%. In addition when applied in the supervised setting our method can freely save up to 60% human annotations without a loss of performance.
****
VISTA-LLAMA: Reducing Hallucination in Video Language Models via Equal Distance to Visual Tokens
Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13151-13160
Recent advances in large video-language models have displayed promising outcomes in video comprehension. Current approaches straightforwardly convert video into language tokens and employ large language models for multi-modal tasks. However this method often leads to the generation of irrelevant content commonly known as "hallucination" as the length of the text increases and the impact of the video diminishes. To address this problem we propose Vista-LLaMA a novel framework that maintains the consistent distance between all visual tokens and any language tokens irrespective of the generated text length. Vista-LLaMA omits relative position encoding when determining attention weights between visual and text tokens retaining the position encoding for text and text tokens. This amplifies the effect of visual tokens on text generation especially when the relative distance is longer between visual and text tokens. The proposed attention mechanism significantly reduces the chance of producing irrelevant text related to the video content. Furthermore we present a sequential visual projector that projects the current video frame into tokens of language space with the assistance of the previous frame. This approach not only captures the temporal relationship within the video but also allows less visual tokens to encompass the entire video. Our approach significantly outperforms various previous methods (e.g. Video-ChatGPT MovieChat) on four challenging open-ended video question answering benchmarks. We reach an accuracy of 60.7 on the zero-shot NExT-QA and 60.5 on the zero-shot MSRVTT-QA setting a new state-of-the-art performance.

****
FSRT: Facial Scene Representation Transformer for Face Reenactment from Factoriz
ed Appearance Head-pose and Facial Expression Features
Andre Rochow, Max Schwarz, Sven Behnke; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7716-7726
The task of face reenactment is to transfer the head motion and facial expressio
ns from a driving video to the appearance of a source image which may be of a di
fferent person (cross-reenactment). Most existing methods are CNN-based and esti
mate optical flow from the source image to the current driving frame which is th
en inpainted and refined to produce the output animation. We propose a transform
er-based encoder for computing a set-latent representation of the source image(s
). We then predict the output color of a query pixel using a transformer-based d
ecoder which is conditioned with keypoints and a facial expression vector extrac
ted from the driving frame. Latent representations of the source person are lear
ned in a self-supervised manner that factorize their appearance head pose and fa
cial expressions. Thus they are perfectly suited for cross-reenactment. In contr
ast to most related work our method naturally extends to multiple source images
and can thus adapt to person-specific facial dynamics. We also propose data augm
entation and regularization schemes that are necessary to prevent overfitting an
d support generalizability of the learned representations. We evaluated our appr
oach in a randomized user study. The results indicate superior performance compa
red to the state-of-the-art in terms of motion transfer quality and temporal con
sistency.
****
Efficient Multitask Dense Predictor via Binarization
Yuzhang Shang, Dan Xu, Gaowen Liu, Ramana Rao Kompella, Yan Yan; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
 pp. 15899-15908
Multi-task learning for dense prediction has emerged as a pivotal area in comput
er vision enabling simultaneous processing of diverse yet interrelated pixel-wis
e prediction tasks. However the substantial computational demands of state-of-th
e-art (SoTA) models often limit their widespread deployment. This paper addresse
s this challenge by introducing network binarization to compress resource-intens
ive multi-task dense predictors. Specifically our goal is to significantly accel
erate multi-task dense prediction models via Binary Neural Networks (BNNs) while
 maintaining and even improving model performance at the same time. To reach thi
s goal we propose a Binary Multi-task Dense Predictor Bi-MTDP and several varian
ts of \bimtdp in which a multi-task dense predictor is constructed via specified
 binarized modules. Our systematical analysis of this predictor reveals that per
formance drop from binarization is primarily caused by severe information degrad
ation. To address this issue we introduce a deep information bottleneck layer th
at enforces representations for downstream tasks satisfying Gaussian distributio
n in forward propagation. Moreover we introduce a knowledge distillation mechani
sm to correct the direction of information flow in backward propagation. Intrigu
ingly one variant of Bi-MTDP outperforms full-precision (FP) multi-task dense pr
ediction SoTAs ARTC (CNN-based) and InvPT (ViT-based). This result indicates tha
t Bi-MTDP is not merely a naive trade-off between performance and efficiency but
 is rather a benefit of the redundant information flow thanks to the multi-task
architecture.
****
TetraSphere: A Neural Descriptor for O(3)-Invariant Point Cloud Analysis
Pavlo Melnyk, Andreas Robinson, Michael Felsberg, Mårten Wadenbäck; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 5620-5630
In many practical applications 3D point cloud analysis requires rotation invaria
nce. In this paper we present a learnable descriptor invariant under 3D rotation
s and reflections i.e. the O(3) actions utilizing the recently introduced steera
ble 3D spherical neurons and vector neurons. Specifically we propose an embeddin
g of the 3D spherical neurons into 4D vector neurons which leverages end-to-end
training of the model. In our approach we perform TetraTransform---an equivarian

t embedding of the 3D input into 4D constructed from the steerable neurons---and extract deeper O(3)-equivariant features using vector neurons. This integration of the TetraTransform into the VN-DGCNN framework termed TetraSphere negligibly increases the number of parameters by less than 0.0002%. TetraSphere sets a new state-of-the-art performance classifying randomly rotated real-world object scans of the challenging subsets of ScanObjectNN. Additionally TetraSphere outperforms all equivariant methods on randomly rotated synthetic data: classifying objects from ModelNet40 and segmenting parts of the ShapeNet shapes. Thus our results reveal the practical value of steerable 3D spherical neurons for learning in 3D Euclidean space. The code is available at https://github.com/pavlo-melnyk/tetrasphere.
****
ZeroRF: Fast Sparse View 360deg Reconstruction with Zero Pretraining
Ruoxi Shi, Xinyue Wei, Cheng Wang, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21114-21124
We present ZeroRF a novel per-scene optimization method addressing the challenge of sparse view 360deg reconstruction in neural field representations. Current breakthroughs like Neural Radiance Fields (NeRF) have demonstrated high-fidelity image synthesis but struggle with sparse input views. Existing methods such as Generalizable NeRFs and per-scene optimization approaches face limitations in data dependency computational cost and generalization across diverse scenarios. To overcome these challenges we propose ZeroRF whose key idea is to integrate a tailored Deep Image Prior into a factorized NeRF representation. Unlike traditional methods ZeroRF parametrizes feature grids with a neural network generator enabling efficient sparse view 360deg reconstruction without any pretraining or additional regularization. Extensive experiments showcase ZeroRF's versatility and superiority in terms of both quality and speed achieving state-of-the-art results on benchmark datasets. ZeroRF's significance extends to applications in 3D content generation and editing. Project page: https://sarahweiii.github.io/zerorf/
****
RCooper: A Real-world Large-scale Dataset for Roadside Cooperative Perception
Ruiyang Hao, Siqi Fan, Yingru Dai, Zhenlin Zhang, Chenxi Li, Yuntian Wang, Haibao Yu, Wenxian Yang, Jirui Yuan, Zaiqing Nie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22347-22357
The value of roadside perception which could extend the boundaries of autonomous driving and traffic management has gradually become more prominent and acknowledged in recent years. However existing roadside perception approaches only focus on the single-infrastructure sensor system which cannot realize a comprehensive understanding of a traffic area because of the limited sensing range and blind spots. Orienting high-quality roadside perception we need Roadside Cooperative Perception (RCooper) to achieve practical area-coverage roadside perception for restricted traffic areas. Rcooper has its own domain-specific challenges but further exploration is hindered due to the lack of datasets. We hence release the first real-world large-scale RCooper dataset to bloom the research on practical roadside cooperative perception including detection and tracking. The manually annotated dataset comprises 50k images and 30k point clouds including two representative traffic scenes (i.e. intersection and corridor). The constructed benchmarks prove the effectiveness of roadside cooperation perception and demonstrate the direction of further research. Codes and dataset can be accessed at: https://github.com/AIR-THU/DAIR-RCooper.
****
TutteNet: Injective 3D Deformations by Composition of 2D Mesh Deformations
Bo Sun, Thibault Groueix, Chen Song, Qixing Huang, Noam Aigerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21378-21389
This work proposes a novel representation of injective deformations of 3D space which overcomes existing limitations of injective methods namely inaccuracy lack of robustness and incompatibility with general learning and optimization frameworks. Our core idea is to reduce the problem to a "deep" composition of multiple 2D mesh-based piecewise-linear maps. Namely we build differentiable layers that

produce mesh deformations through Tutte's embedding (guaranteed to be injective in 2D) and compose these layers over different planes to create complex 3D injective deformations of the 3D volume. We show our method provides the ability to ef?ciently and accurately optimize and learn complex deformations outperforming other injective approaches. As a main application we produce complex and artifact-free NeRF and SDF deformations.
****
WANDR: Intention-guided Human Motion Generation
Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 927-936
Synthesizing natural human motions that enable a 3D human avatar to walk and reach for arbitrary goals in 3D space remains an unsolved problem with many applications. Existing methods (data-driven or using reinforcement learning) are limited in terms of generalization and motion naturalness. A primary obstacle is the scarcity of training data that combines locomotion with goal reaching. To address this we introduce WANDR a data-driven model that takes an avatar's initial pose and a goal's 3D position and generates natural human motions that place the end effector (wrist) on the goal location. To solve this we introduce novel intention features that drive rich goal-oriented movement. Intention guides the agent to the goal and interactively adapts the generation to novel situations without needing to define sub-goals or the entire motion path. Crucially intention allows training on datasets that have goal-oriented motions as well as those that do not. WANDR is a conditional Variational Auto-Encoder (c-VAE) which we train using the AMASS and CIRCLE datasets. We evaluate our method extensively and demonstrate its ability to generate natural and long-term motions that reach 3D goals and generalize to unseen goal locations. Our models and code are available for research purposes at wandr.is.tue.mpg.de
****
Jointly Training and Pruning CNNs via Learnable Agent Guidance and Alignment
Alireza Ganjdanesh, Shangqian Gao, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16058-16069
Structural model pruning is a prominent approach used for reducing the computational cost of Convolutional Neural Networks (CNNs) before their deployment on resource-constrained devices. Yet the majority of proposed ideas require a pretrained model before pruning which is costly to secure. In this paper we propose a novel structural pruning approach to jointly learn the weights and structurally prune architectures of CNN models. The core element of our method is a Reinforcement Learning (RL) agent whose actions determine the pruning ratios of the CNN model's layers and the resulting model's accuracy serves as its reward. We conduct the joint training and pruning by iteratively training the model's weights and the agent's policy and we regularize the model's weights to align with the selected structure by the agent. The evolving model's weights result in a dynamic reward function for the agent which prevents using prominent episodic RL methods with stationary environment assumption for our purpose. We address this challenge by designing a mechanism to model the complex changing dynamics of the reward function and provide a representation of it to the RL agent. To do so we take a learnable embedding for each training epoch and employ a recurrent model to calculate a representation of the changing environment. We train the recurrent model and embeddings using a decoder model to reconstruct observed rewards. Such a design empowers our agent to effectively leverage episodic observations along with the environment representations to learn a proper policy to determine performant sub-networks of the CNN model. Our extensive experiments on CIFAR-10 and ImageNet using ResNets and MobileNets demonstrate the effectiveness of our method.
****
Estimating Noisy Class Posterior with Part-level Labels for Noisy Label Learning
Rui Zhao, Bin Shi, Jianfei Ruan, Tianze Pan, Bo Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22809-22819
In noisy label learning estimating noisy class posteriors plays a fundamental ro

le for developing consistent classifiers as it forms the basis for estimating clean class posteriors and the transition matrix. Existing methods typically learn noisy class posteriors by training a classification model with noisy labels. However when labels are incorrect these models may be misled to overemphasize the feature parts that do not reflect the instance characteristics resulting in significant errors in estimating noisy class posteriors. To address this issue this paper proposes to augment the supervised information with part-level labels encouraging the model to focus on and integrate richer information from various parts. Specifically our method first partitions features into distinct parts by cropping instances yielding part-level labels associated with these various parts. Subsequently we introduce a novel single-to-multiple transition matrix to model the relationship between the noisy and part-level labels which incorporates part-level labels into a classifier-consistent framework. Utilizing this framework with part-level labels we can learn the noisy class posteriors more precisely by guiding the model to integrate information from various parts ultimately improving the classification performance. Our method is theoretically sound while experiments show that it is empirically effective in synthetic and real-world noisy benchmarks.
****
Leveraging Vision-Language Models for Improving Domain Generalization in Image Classification
Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23922-23932
Vision-Language Models (VLMs) such as CLIP are trained on large amounts of image-text pairs resulting in remarkable generalization across several data distributions. However in several cases their expensive training and data collection/curation costs do not justify the end application. This motivates a vendor-client paradigm where a vendor trains a large-scale VLM and grants only input-output access to clients on a pay-per-query basis in a black-box setting. The client aims to minimize inference cost by distilling the VLM to a student model using the limited available task-specific data and further deploying this student model in the downstream application. While naive distillation largely improves the In-Domain (ID) accuracy of the student it fails to transfer the superior out-of-distribution (OOD) generalization of the VLM teacher using the limited available labeled images. To mitigate this we propose Vision-Language to Vision - Align Distill Predict (VL2V-ADiP) which first aligns the vision and language modalities of the teacher model with the vision modality of a pre-trained student model and further distills the aligned VLM representations to the student. This maximally retains the pre-trained features of the student while also incorporating the rich representations of the VLM image encoder and the superior generalization of the text embeddings. The proposed approach achieves state-of-the-art results on the standard Domain Generalization benchmarks in a black-box teacher setting as well as a white-box setting where the weights of the VLM are accessible.
****
Diffusion-EDFs: Bi-equivariant Denoising Generative Modeling on SE(3) for Visual Robotic Manipulation
Hyunwoo Ryu, Jiwoo Kim, Hyunseok An, Junwoo Chang, Joohwan Seo, Taehan Kim, Yubin Kim, Chaewon Hwang, Jongeun Choi, Roberto Horowitz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18007-18018
Diffusion generative modeling has become a promising approach for learning robotic manipulation tasks from stochastic human demonstrations. In this paper we present Diffusion-EDFs a novel SE(3)-equivariant diffusion-based approach for visual robotic manipulation tasks. We show that our proposed method achieves remarkable data efficiency requiring only 5 to 10 human demonstrations for effective end-to-end training in less than an hour. Furthermore our benchmark experiments demonstrate that our approach has superior generalizability and robustness compared to state-of-the-art methods. Lastly we validate our methods with real hardware experiments.

****
Prompt Learning via Meta-Regularization
Jinyoung Park, Juyeon Ko, Hyunwoo J. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26940-26950
Pre-trained vision-language models have shown impressive success on various computer vision tasks with their zero-shot generalizability. Recently prompt learning approaches have been explored to efficiently and effectively adapt the vision-language models to a variety of downstream tasks. However most existing prompt learning methods suffer from task overfitting since the general knowledge of the pre-trained vision language models is forgotten while the prompts are finetuned on a small data set from a specific target task. To address this issue we propose a Prompt Meta-Regularization (ProMetaR) to improve the generalizability of prompt learning for vision-language models. Specifically ProMetaR meta-learns both the regularizer and the soft prompts to harness the task-specific knowledge from the downstream tasks and task-agnostic general knowledge from the vision-language models. Further ProMetaR augments the task to generate multiple virtual tasks to alleviate the meta-overfitting. In addition we provide the analysis to comprehend how ProMetaR improves the generalizability of prompt tuning in the perspective of the gradient alignment. Our extensive experiments demonstrate that our ProMetaR improves the generalizability of conventional prompt learning methods under base-to-base/base-to-new and domain generalization settings. The code of ProMetaR is available at https://github.com/mlvlab/ProMetaR.
****
Contrasting Intra-Modal and Ranking Cross-Modal Hard Negatives to Enhance Visio-Linguistic Compositional Understanding
Le Zhang, Rabiul Awal, Aishwarya Agrawal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13774-13784
Vision-Language Models (VLMs) such as CLIP exhibit strong image-text comprehension abilities facilitating advances in several downstream tasks such as zero-shot image classification image-text retrieval and text-to-image generation. However the compositional reasoning abilities of existing VLMs remains subpar. The root of this limitation lies in the inadequate alignment between the images and captions in the pretraining datasets. Additionally the current contrastive learning objective fails to focus on fine-grained grounding components like relations actions and attributes resulting in "bag-of-words" representations. We introduce a simple and effective method to improve compositional reasoning in VLMs. Our method better leverages available datasets by refining and expanding the standard image-text contrastive learning framework. Our approach does not require specific annotations and does not incur extra parameters. When integrated with CLIP our technique yields notable improvement over state-of-the-art baselines across five vision-language compositional benchmarks.
****
CMA: A Chromaticity Map Adapter for Robust Detection of Screen-Recapture Document Images
Changsheng Chen, Liangwei Lin, Yongqi Chen, Bin Li, Jishen Zeng, Jiwu Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15577-15586
The rebroadcasting of screen-recaptured document images introduces a significant risk to the confidential documents processed in government departments and commercial companies. However detecting recaptured document images subjected to distortions from online social networks (OSNs) is challenging since the common forensics cues such as moir ?e pattern are weakened during transmission. In this work we first devise a pixel-level distortion model of the screen-recaptured document image to identify the robust features of color artifacts. Then we extract a chromaticity map from the recaptured image to highlight the presence of color artifacts even under low-quality samples. Based on the prior understanding we design a chromaticity map adapter (CMA) to efficiently extract the chromaticity map and feed it into the transformer backbone as multi-modal prompt tokens. To evaluate the performance of the proposed method we collect a recaptured office document image dataset with over 10K diverse samples. Experimental results demonstrate t

hat the proposed CMA method outperforms a SOTA approach (with RGB modality only) reducing the average EER from 26.82% to 16.78%. Robustness evaluation shows that our method achieves 0.8688 and 0.7554 AUCs under samples with JPEG compression (QF=70) and resolution as low as 534x503 pixels.
****
Embodied Multi-Modal Agent trained by an LLM from a Parallel TextWorld
Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, Yuhui Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26275-26285
While large language models (LLMs) excel in a simulated world of texts they struggle to interact with the more realistic world without perceptions of other modalities such as visual or audio signals. Although vision-language models (VLMs) integrate LLM modules (1) aligned with static image features and (2) may possess prior knowledge of world dynamics (as demonstrated in the text world) they have not been trained in an embodied visual world and thus cannot align with its dynamics. On the other hand training an embodied agent in a noisy visual world without expert guidance is often challenging and inefficient. In this paper we train a VLM agent living in a visual world using an LLM agent excelling in a parallel text world. Specifically we distill LLM's reflection outcomes (improved actions by analyzing mistakes) in a text world's tasks to finetune the VLM on the same tasks of the visual world resulting in an Embodied Multi-Modal Agent (EMMA) quickly adapting to the visual world dynamics. Such cross-modality imitation learning between the two parallel worlds is achieved by a novel DAgger-DPO algorithm enabling EMMA to generalize to a broad scope of new tasks without any further guidance from the LLM expert. Extensive evaluations on the ALFWorld benchmark's diverse tasks highlight EMMA's superior performance to SOTA VLM-based agents e.g. 20%-70% improvement in the success rate.
****
VA3: Virtually Assured Amplification Attack on Probabilistic Copyright Protection for Text-to-Image Generative Models
Xiang Li, Qianli Shen, Kenji Kawaguchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12363-12373
The booming use of text-to-image generative models has raised concerns about their high risk of producing copyright-infringing content. While probabilistic copyright protection methods provide a probabilistic guarantee against such infringement in this paper we introduce Virtually Assured Amplification Attack (VA3) a novel online attack framework that exposes the vulnerabilities of these protection mechanisms. The proposed framework significantly amplifies the probability of generating infringing content on the sustained interactions with generative models and a non-trivial lower-bound on the success probability of each engagement. Our theoretical and experimental results demonstrate the effectiveness of our approach under various scenarios.These findings highlight the potential risk of implementing probabilistic copyright protection in practical applications of text-to-image generative models. Code is available at https://github.com/South7X/VA3.
****
Point-VOS: Pointing Up Video Object Segmentation
Sabarinath Mahadevan, Idil Esen Zulfikar, Paul Voigtlaender, Bastian Leibe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22217-22226
Current state-of-the-art Video Object Segmentation (VOS) methods rely on dense per-object mask annotations both during training and testing. This requires time-consuming and costly video annotation mechanisms. We propose a novel Point-VOS task with a spatio-temporally sparse point-wise annotation scheme that substantially reduces the annotation effort. We apply our annotation scheme to two large-scale video datasets with text descriptions and annotate over 19M points across 133K objects in 32K videos. Based on our annotations we propose a new Point-VOS benchmark and a corresponding point-based training mechanism which we use to establish strong baseline results. We show that existing VOS methods can easily be adapted to leverage our point annotations during training and can achieve results close to the fully-supervised performance when trained on pseudo-masks generate

d from these points. In addition we show that our data can be used to improve models that connect vision and language by evaluating it on the Video Narrative Grounding (VNG) task. We will make our code and annotations available at https://pointvos.github.io.
****
Intriguing Properties of Diffusion Models: An Empirical Study of the Natural Attack Capability in Text-to-Image Generative Models
Takami Sato, Justin Yue, Nanze Chen, Ningfei Wang, Qi Alfred Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24635-24644
Denoising probabilistic diffusion models have shown breakthrough performance to generate more photo-realistic images or human-level illustrations than the prior models such as GANs. This high image-generation capability has stimulated the creation of many downstream applications in various areas. However we find that this technology is actually a double-edged sword: We identify a new type of attack called the Natural Denoising Diffusion (NDD) attack based on the finding that state-of-the-art deep neural network (DNN) models still hold their prediction even if we intentionally remove their robust features which are essential to the human visual system (HVS) through text prompts. The NDD attack shows a significantly high capability to generate low-cost model-agnostic and transferable adversarial attacks by exploiting the natural attack capability in diffusion models. To systematically evaluate the risk of the NDD attack we perform a large-scale empirical study with our newly created dataset the Natural Denoising Diffusion Attack (NDDA) dataset. We evaluate the natural attack capability by answering 6 research questions. Through a user study we find that it can achieve an 88% detection rate while being stealthy to 93% of human subjects; we also find that the non-robust features embedded by diffusion models contribute to the natural attack capability. To confirm the model-agnostic and transferable attack capability we perform the NDD attack against the Tesla Model 3 and find that 73% of the physically printed attacks can be detected as stop signs. Our hope is that the study and dataset can help our community be aware of the risks in diffusion models and facilitate further research toward robust DNN models.
****
GroupContrast: Semantic-aware Self-supervised Representation Learning for 3D Understanding
Chengyao Wang, Li Jiang, Xiaoyang Wu, Zhuotao Tian, Bohao Peng, Hengshuang Zhao, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4917-4928
Self-supervised 3D representation learning aims to learn effective representations from large-scale unlabeled point clouds. Most existing approaches adopt point discrimination as the pretext task which assigns matched points in two distinct views as positive pairs and unmatched points as negative pairs. However this approach often results in semantically identical points having dissimilar representations leading to a high number of false negatives and introducing a semantic conflict problem. To address this issue we propose GroupContrast a novel approach that combines segment grouping and semantic-aware contrastive learning. Segment grouping partitions points into semantically meaningful regions which enhances semantic coherence and provides semantic guidance for the subsequent contrastive representation learning. Semantic-aware contrastive learning augments the semantic information extracted from segment grouping and helps to alleviate the issue of semantic conflict. We conducted extensive experiments on multiple 3D scene understanding tasks. The results demonstrate that GroupContrast learns semantically meaningful representations and achieves promising transfer learning performance.
****
HouseCat6D - A Large-Scale Multi-Modal Category Level 6D Object Perception Dataset with Household Objects in Realistic Scenarios
HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, Daniel Roth, Nassir Navab, Benjamin Busam; Proceedings of the IEEE/CVF Conference on

 Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22498-22508
Estimating 6D object poses is a major challenge in 3D computer vision. Building on successful instance-level approaches research is shifting towards category-level pose estimation for practical applications. Current category-level datasets however fall short in annotation quality and pose variety. Addressing this we introduce HouseCat6D a new category-level 6D pose dataset. It features 1) multi-modality with Polarimetric RGB and Depth (RGBD+P) 2) encompasses 194 diverse objects across 10 household categories including two photometrically challenging ones and 3) provides high-quality pose annotations with an error range of only 1.35 mm to 1.74 mm. The dataset also includes 4) 41 large-scale scenes with comprehensive viewpoint and occlusion coverage 5) a checkerboard-free environment and 6. dense 6D parallel-jaw robotic grasp annotations. Additionally we present benchmark results for leading category-level pose estimation networks.
****
Privacy-Preserving Face Recognition Using Trainable Feature Subtraction
Yuxi Mi, Zhizhou Zhong, Yuge Huang, Jiazhen Ji, Jianqing Xu, Jun Wang, Shaoming Wang, Shouhong Ding, Shuigeng Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 297-307
The widespread adoption of face recognition has led to increasing privacy concerns as unauthorized access to face images can expose sensitive personal information. This paper explores face image protection against viewing and recovery attacks. Inspired by image compression we propose creating a visually uninformative face image through feature subtraction between an original face and its model-produced regeneration. Recognizable identity features within the image are encouraged by co-training a recognition model on its high-dimensional feature representation. To enhance privacy the high-dimensional representation is crafted through random channel shuffling resulting in randomized recognizable images devoid of attacker-leverageable texture details. We distill our methodologies into a novel privacy-preserving face recognition method MinusFace. Experiments demonstrate its high recognition accuracy and effective privacy protection. Its code is available at https://github.com/Tencent/TFace.
****
Towards Co-Evaluation of Cameras HDR and Algorithms for Industrial-Grade 6DoF Pose Estimation
Agastya Kalra, Guy Stoppi, Dmitrii Marin, Vage Taamazyan, Aarrushi Shandilya, Rishav Agarwal, Anton Boykov, Tze Hao Chong, Michael Stark; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22691-22701
6DoF Pose estimation has been gaining increased importance in vision for over a decade however it does not yet meet the reliability and accuracy standards for mass deployment in industrial robotics. To this effect we present the Industrial Plenoptic Dataset (IPD): the first dataset for the co-evaluation of cameras HDR and algorithms targeted at reliable high-accuracy industrial automation. Specifically we capture 2300 physical scenes of 20 industrial parts covering a 1mx1mx0.5m working volume resulting in over 100000 distinct object views. Each scene is captured with 13 well-calibrated multi-modal cameras including polarization and high-resolution structured light. In terms of lighting we capture each scene at 4 exposures and in 3 challenging lighting conditions ranging from 100 lux to 100000 lux. We also present validate and analyze robot consistency an evaluation method targeted at scalable high accuracy evaluation. We hope that vision systems that succeed on this dataset will have direct industry impact. The dataset and evaluation code are available at https://github.com/intrinsic-ai/ipd.
****
Learning Visual Prompt for Gait Recognition
Kang Ma, Ying Fu, Chunshui Cao, Saihui Hou, Yongzhen Huang, Dezhi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 593-603
Gait a prevalent and complex form of human motion plays a significant role in the field of long-range pedestrian retrieval due to the unique characteristics inherent in individual motion patterns. However gait recognition in real-world scen

arios is challenging due to the limitations of capturing comprehensive cross-vie
wing and cross-clothing data. Additionally distractors such as occlusions direct
ional changes and lingering movements further complicate the problem. The widesp
read application of deep learning techniques has led to the development of vario
us potential gait recognition methods. However these methods utilize convolution
al networks to extract shared information across different views and attire cond
itions. Once trained the parameters and non-linear function become constrained t
o fixed patterns limiting their adaptability to various distractors in real-worl
d scenarios. In this paper we present a unified gait recognition framework to ex
tract global motion patterns and develop a novel dynamic transformer to generate
 representative gait features. Specifically we develop a trainable part-based pr
ompt pool with numerous key-value pairs that can dynamically select prompt templ
ates to incorporate into the gait sequence thereby providing task-relevant share
d knowledge information. Furthermore we specifically design dynamic attention to
 extract robust motion patterns and address the length generalization issue. Ext
ensive experiments on four widely recognized gait datasets i.e. Gait3D GREW OUMV
LP and CASIA-B reveal that the proposed method yields substantial improvements c
ompared to current state-of-the-art approaches.
****
MLP Can Be A Good Transformer Learner
Sihao Lin, Pumeng Lyu, Dongrui Liu, Tao Tang, Xiaodan Liang, Andy Song, Xiaojun
Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2024, pp. 19489-19498
Self-attention mechanism is the key of the Transformer but often criticized for
its computation demands. Previous token pruning works motivate their methods fro
m the view of computation redundancy but still need to load the full network and
 require same memory costs. This paper introduces a novel strategy that simplifi
es vision transformers and reduces computational load through the selective remo
val of non-essential attention layers guided by entropy considerations. We ident
ify that regarding the attention layer in bottom blocks their subsequent MLP lay
ers i.e. two feed-forward layers can elicit the same entropy quantity. Meanwhile
 the accompanied MLPs are under-exploited since they exhibit smaller feature ent
ropy compared to those MLPs in the top blocks. Therefore we propose to integrate
 the uninformative attention layers into their subsequent counterparts by degene
rating them into identical mapping yielding only MLP in certain transformer bloc
ks. Experimental results on ImageNet-1k show that the proposed method can remove
 40% attention layer of DeiT-B improving throughput and memory bound without per
formance compromise.
****
GraphDreamer: Compositional 3D Scene Synthesis from Scene Graphs
Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, Bernhard Schölkopf; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2024, pp. 21295-21304
As pretrained text-to-image diffusion models become increasingly powerful recent
 efforts have been made to distill knowledge from these text-to-image pretrained
 models for optimizing a text-guided 3D model. Most of the existing methods gene
rate a holistic 3D model from a plain text input. This can be problematic when t
he text describes a complex scene with multiple objects because the vectorized t
ext embeddings are inherently unable to capture a complex description with multi
ple entities and relationships. Holistic 3D modeling of the entire scene further
 prevents accurate grounding of text entities and concepts. To address this limi
tation we propose GraphDreamer a novel framework to generate compositional 3D sc
enes from scene graphs where objects are represented as nodes and their interact
ions as edges. By exploiting node and edge information in scene graphs our metho
d makes better use of the pretrained text-to-image diffusion model and is able t
o fully disentangle different objects without image-level supervision. To facili
tate modeling of object-wise relationships we use signed distance fields as repr
esentation and impose a constraint to avoid inter-penetration of objects. To avo
id manual scene graph creation we design a text prompt for ChatGPT to generate s
cene graphs based on text inputs. We conduct both qualitative and quantitative e

xperiments to validate the effectiveness of GraphDreamer in generating high-fide lity compositional 3D scenes with disentangled object entities.
****
Visual-Augmented Dynamic Semantic Prototype for Generative Zero-Shot Learning
Wenjin Hou, Shiming Chen, Shuhuang Chen, Ziming Hong, Yan Wang, Xuetao Feng, Sal man Khan, Fahad Shahbaz Khan, Xinge You; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23627-23637
Generative Zero-shot learning (ZSL) learns a generator to synthesize visual samp les for unseen classes which is an effective way to advance ZSL. However existin g generative methods rely on the conditions of Gaussian noise and the predefined semantic prototype which limit the generator only optimized on specific seen cl asses rather than characterizing each visual instance resulting in poor generali zations (e.g. overfitting to seen classes). To address this issue we propose a n ovel Visual-Augmented Dynamic Semantic prototype method (termed VADS) to boost t he generator to learn accurate semantic-visual mapping by fully exploiting the v isual-augmented knowledge into semantic conditions. In detail VADS consists of t wo modules: (1) Visual-aware Domain Knowledge Learning module (VDKL) learns the local bias and global prior of the visual features (referred to as domain visual knowledge) which replace pure Gaussian noise to provide richer prior noise info rmation; (2) Vision-Oriented Semantic Updation module (VOSU) updates the semanti c prototype according to the visual representations of the samples. Ultimately w e concatenate their output as a dynamic semantic prototype which serves as the c ondition of the generator. Extensive experiments demonstrate that our VADS achie ves superior CZSL and GZSL performances on three prominent datasets and outperfo rms other state-of-the-art methods with averaging increases by 6.4% 5.9% and 4.2 % on SUN CUB and AWA2 respectively.
****
Dynamic Prompt Optimizing for Text-to-Image Generation
Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, Qing Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 024, pp. 26627-26636
Text-to-image generative models specifically those based on diffusion models lik e Imagen and Stable Diffusion have made substantial advancements. Recently there has been a surge of interest in the delicate refinement of text prompts. Users assign weights or alter the injection time steps of certain words in the text pr ompts to improve the quality of generated images. However the success of fine-co ntrol prompts depends on the accuracy of the text prompts and the careful select ion of weights and time steps which requires significant manual intervention. To address this we introduce the Prompt Auto-Editing (PAE) method. Besides refinin g the original prompts for image generation we further employ an online reinforc ement learning strategy to explore the weights and injection time steps of each word leading to the dynamic fine-control prompts. The reward function during tra ining encourages the model to consider aesthetic score semantic consistency and user preferences. Experimental results demonstrate that our proposed method effe ctively improves the original prompts generating visually more appealing images while maintaining semantic alignment. Code is available at \href https://github. com/Mowenyii/PAE this https URL .
****
SC-GS: Sparse-Controlled Gaussian Splatting for Editable Dynamic Scenes
Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2024, pp. 4220-4230
Novel view synthesis for dynamic scenes is still a challenging problem in comput er vision and graphics. Recently Gaussian splatting has emerged as a robust tech nique to represent static scenes and enable high-quality and real-time novel vie w synthesis. Building upon this technique we propose a new representation that e xplicitly decomposes the motion and appearance of dynamic scenes into sparse con trol points and dense Gaussians respectively. Our key idea is to use sparse cont rol points significantly fewer in number than the Gaussians to learn compact 6 D oF transformation bases which can be locally interpolated through learned interp

olation weights to yield the motion field of 3D Gaussians. We employ a deformati
on MLP to predict time-varying 6 DoF transformations for each control point whic
h reduces learning complexities enhances learning abilities and facilitates obta
ining temporal and spatial coherent motion patterns. Then we jointly learn the 3
D Gaussians the canonical space locations of control points and the deformation
MLP to reconstruct the appearance geometry and dynamics of 3D scenes. During lea
rning the location and number of control points are adaptively adjusted to accom
modate varying motion complexities in different regions and an ARAP loss followi
ng the principle of as rigid as possible is developed to enforce spatial continu
ity and local rigidity of learned motions. Finally thanks to the explicit sparse
motion representation and its decomposition from appearance our method can enab
le user-controlled motion editing while retaining high-fidelity appearances. Ext
ensive experiments demonstrate that our approach outperforms existing approaches
on novel view synthesis with a high rendering speed and enables novel appearanc
e-preserved motion editing applications.
****
360Loc: A Dataset and Benchmark for Omnidirectional Visual Localization with Cro
ss-device Queries
Huajian Huang, Changkun Liu, Yipeng Zhu, Hui Cheng, Tristan Braud, Sai-Kit Yeung
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2024, pp. 22314-22324
Portable 360^\circ cameras are becoming a cheap and efficient tool to establish
large visual databases. By capturing omnidirectional views of a scene these came
ras could expedite building environment models that are essential for visual loc
alization. However such an advantage is often overlooked due to the lack of valu
able datasets. This paper introduces a new benchmark dataset 360Loc composed of
360^\circ images with ground truth poses for visual localization. We present a p
ractical implementation of 360^\circ mapping combining 360^\circ images with lid
ar data to generate the ground truth 6DoF poses. 360Loc is the first dataset and
benchmark that explores the challenge of cross-device visual positioning involv
ing 360^\circ reference frames and query frames from pinhole ultra-wide FoV fish
eye and 360^\circ cameras. We propose a virtual camera approach to generate lowe
r-FoV query frames from 360^\circ images which ensures a fair comparison of perf
ormance among different query types in visual localization tasks. We also extend
this virtual camera approach to feature matching-based and pose regression-base
d methods to alleviate the performance loss caused by the cross-device domain ga
p and evaluate its effectiveness against state-of-the-art baselines. We demonstr
ate that omnidirectional visual localization is more robust in challenging large
-scale scenes with symmetries and repetitive structures. These results provide n
ew insights into 360-camera mapping and omnidirectional visual localization with
cross-device queries. Project Page and dataset: https://huajianup.github.io/res
earch/360Loc/.
****
Domain Gap Embeddings for Generative Dataset Augmentation
Yinong Oliver Wang, Younjoon Chung, Chen Henry Wu, Fernando De la Torre; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 28684-28694
The performance of deep learning models is intrinsically tied to the quality vol
ume and relevance of their training data. Gathering ample data for production sc
enarios often demands significant time and resources. Among various strategies d
ata augmentation circumvents exhaustive data collection by generating new data p
oints from existing ones. However traditional augmentation techniques can be les
s effective amidst a shift in training and testing distributions. This paper exp
lores the potential of synthetic data by leveraging large pre-trained models for
data augmentation especially when confronted with distribution shifts. Although
recent advancements in generative models have enabled several prior works in cr
oss-distribution data generation they require model fine-tuning and a complex se
tup. To bypass these shortcomings we introduce Domain Gap Embeddings (DoGE) a pl
ug-and-play semantic data augmentation framework in a cross-distribution few-sho
t setting. Our method extracts disparities between source and desired data distr

ibutions in a latent form and subsequently steers a generative process to supplement the training set with endless diverse synthetic samples. Our evaluations conducted on a subpopulation shift and three domain adaptation scenarios under a few-shot paradigm reveal that our versatile method improves performance across tasks without needing hands-on intervention or intricate fine-tuning. DoGE paves the way to effortlessly generate realistic controllable synthetic datasets following the test distributions bolstering real-world efficacy for downstream task models.
****
Geometrically-driven Aggregation for Zero-shot 3D Point Cloud Understanding
Guofeng Mei, Luigi Riz, Yiming Wang, Fabio Poiesi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27896-27905
Zero-shot 3D point cloud understanding can be achieved via 2D Vision-Language Models (VLMs). Existing strategies directly map VLM representations from 2D pixels of rendered or captured views to 3D points overlooking the inherent and expressible point cloud geometric structure. Geometrically similar or close regions can be exploited for bolstering point cloud understanding as they are likely to share semantic information. To this end we introduce the first training-free aggregation technique that leverages the point cloud's 3D geometric structure to improve the quality of the transferred VLM representation. Our approach operates iteratively performing local-to-global aggregation based on geometric and semantic point-level reasoning. We benchmark our approach on three downstream tasks including classification part segmentation and semantic segmentation with a variety of datasets representing both synthetic/real-world and indoor/outdoor scenarios. Our approach achieves new state-of-the-art results in all benchmarks.
****
Learning to Rank Patches for Unbiased Image Redundancy Reduction
Yang Luo, Zhineng Chen, Peng Zhou, Zuxuan Wu, Xieping Gao, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22831-22840
Images suffer from heavy spatial redundancy because pixels in neighboring regions are spatially correlated. Existing approaches strive to overcome this limitation by reducing less meaningful image regions. However current leading methods rely on supervisory signals. They may compel models to preserve content that aligns with labeled categories and discard content belonging to unlabeled categories. This categorical inductive bias makes these methods less effective in real-world scenarios. To address this issue we propose a self-supervised framework for image redundancy reduction called Learning to Rank Patches (LTRP). We observe that image reconstruction of masked image modeling models is sensitive to the removal of visible patches when the masking ratio is high (e.g. 90%). Building upon it we implement LTRP via two steps: inferring the semantic density score of each patch by quantifying variation between reconstructions with and without this patch and learning to rank the patches with the pseudo score. The entire process is self-supervised thus getting out of the dilemma of categorical inductive bias. We design extensive experiments on different datasets and tasks. The results demonstrate that LTRP outperforms both supervised and other self-supervised methods due to the fair assessment of image content.
****
Going Beyond Multi-Task Dense Prediction with Synergy Embedding Models
Huimin Huang, Yawen Huang, Lanfen Lin, Ruofeng Tong, Yen-Wei Chen, Hao Zheng, Yuexiang Li, Yefeng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28181-28190
Multi-task visual scene understanding aims to leverage the relationships among a set of correlated tasks which are solved simultaneously by embedding them within a uni- fied network. However most existing methods give rise to two primary concerns from a task-level perspective: (1) the lack of task-independent correspondences for distinct tasks and (2) the neglect of explicit task-consensual dependencies among various tasks. To address these issues we propose a novel synergy embedding models (SEM) which goes be- yond multi-task dense prediction by leverag

ing two innova- tive designs: the intra-task hierarchy-adaptive module and the i
nter-task EM-interactive module. Specifically the con- structed intra-task modul
e incorporates hierarchy-adaptive keys from multiple stages enabling the efficie
nt learning of specialized visual patterns with an optimal trade-off. In ad- dit
ion the developed inter-task module learns interactions from a compact set of mu
tual bases among various tasks benefiting from the expectation maximization (EM)
 algo- rithm. Extensive empirical evidence from two public bench- marks NYUD-v2
and PASCAL-Context demonstrates that SEM consistently outperforms state-of-the-a
rt approaches across a range of metrics.
****
Disentangled Pre-training for Human-Object Interaction Detection
Zhuolong Li, Xingao Li, Changxing Ding, Xiangmin Xu; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28191-2
8201
Detecting human-object interaction (HOI) has long been limited by the amount of
supervised data available. Recent approaches address this issue by pre-training
according to pseudo-labels which align object regions with HOI triplets parsed f
rom image captions. However pseudo-labeling is tricky and noisy making HOI pre-t
raining a complex process. Therefore we propose an efficient disentangled pre-tr
aining method for HOI detection (DP-HOI) to address this problem. First DP-HOI u
tilizes object detection and action recognition datasets to pre-train the detect
ion and interaction decoder layers respectively. Then we arrange these decoder l
ayers so that the pre-training architecture is consistent with the downstream HO
I detection task. This facilitates efficient knowledge transfer. Specifically th
e detection decoder identifies reliable human instances in each action recogniti
on dataset image generates one corresponding query and feeds it into the interac
tion decoder for verb classification. Next we combine the human instance verb pr
edictions in the same image and impose image-level supervision. The DP-HOI struc
ture can be easily adapted to the HOI detection task enabling effective model pa
rameter initialization. Therefore it significantly enhances the performance of e
xisting HOI detection models on a broad range of rare categories. The code and p
re-trained weight are available at https://github.com/xingaoli/DP-HOI.
****
Light the Night: A Multi-Condition Diffusion Framework for Unpaired Low-Light En
hancement in Autonomous Driving
Jinlong Li, Baolu Li, Zhengzhong Tu, Xinyu Liu, Qing Guo, Felix Juefei-Xu, Runsh
eng Xu, Hongkai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2024, pp. 15205-15215
Vision-centric perception systems for autonomous driving have gained considerabl
e attention recently due to their cost-effectiveness and scalability especially
compared to LiDAR-based systems. However these systems often struggle in low-lig
ht conditions potentially compromising their performance and safety. To address
this our paper introduces LightDiff a domain-tailored framework designed to enha
nce the low-light image quality for autonomous driving applications. Specificall
y we employ a multi-condition controlled diffusion model. LightDiff works withou
t any human-collected paired data leveraging a dynamic data degradation process
instead. It incorporates a novel multi-condition adapter that adaptively control
s the input weights from different modalities including depth maps RGB images an
d text captions to effectively illuminate dark scenes while maintaining context
consistency. Furthermore to align the enhanced images with the detection model's
 knowledge LightDiff employs perception-specific scores as rewards to guide the
diffusion training process through reinforcement learning. Extensive experiments
 on the nuScenes datasets demonstrate that LightDiff can significantly improve t
he performance of several state-of-the-art 3D detectors in night-time conditions
 while achieving high visual quality scores highlighting its potential to safegu
ard autonomous driving.
****
MetaCloak: Preventing Unauthorized Subject-driven Text-to-image Diffusion-based
Synthesis via Meta-learning
Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, Lichao Sun; Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 24219-24228
Text-to-image diffusion models allow seamless generation of personalized images
from scant reference photos. Yet these tools in the wrong hands can fabricate mi
sleading or harmful content endangering individuals. To address this problem exi
sting poisoning-based approaches perturb user images in an imperceptible way to
render them "unlearnable" from malicious uses. We identify two limitations of th
ese defending approaches: i) sub-optimal due to the hand-crafted heuristics for
solving the intractable bilevel optimization and ii) lack of robustness against
simple data transformations like Gaussian filtering. To solve these challenges w
e propose MetaCloak which solves the bi-level poisoning problem with a meta-lear
ning framework with an additional transformation sampling process to craft trans
ferable and robust perturbation. Specifically we employ a pool of surrogate diff
usion models to craft transferable and model-agnostic perturbation. Furthermore
by incorporating an additional transformation process we design a simple denoisi
ng-error maximization loss that is sufficient for causing transformation-robust
semantic distortion and degradation in a personalized generation. Extensive expe
riments on the VGGFace2 and CelebA-HQ datasets show that MetaCloak outperforms e
xisting approaches. Notably MetaCloak can successfully fool online training serv
ices like Replicate in a black-box manner demonstrating the effectiveness of Met
aCloak in real-world scenarios.
****
Neural Modes: Self-supervised Learning of Nonlinear Modal Subspaces
Jiahong Wang, Yinwei Du, Stelian Coros, Bernhard Thomaszewski; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 23158-23167
We propose a self-supervised approach for learning physics-based subspaces for r
eal-time simulation. Existing learning-based methods construct subspaces by appr
oximating pre-defined simulation data in a purely geometric way. However this ap
proach tends to produce high-energy configurations leads to entangled latent spa
ce dimensions and generalizes poorly beyond the training set. To overcome these
limitations we propose a self-supervised approach that directly minimizes the sy
stem's mechanical energy during training. We show that our method leads to learn
ed subspaces that reflect physical equilibrium constraints resolve overfitting i
ssues of previous methods and offer interpretable latent space parameters.
****
How to Train Neural Field Representations: A Comprehensive Study and Benchmark
Samuele Papa, Riccardo Valperga, David Knigge, Miltiadis Kofinas, Phillip Lippe,
 Jan-Jakob Sonke, Efstratios Gavves; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2024, pp. 22616-22625
Neural fields (NeFs) have recently emerged as a versatile method for modeling si
gnals of various modalities including images shapes and scenes. Subsequently a n
umber of works have explored the use of NeFs as representations for downstream t
asks e.g. classifying an image based on the parameters of a NeF that has been fi
t to it. However the impact of the NeF hyperparameters on their quality as downs
tream representation is scarcely understood and remains largely unexplored. This
 is in part caused by the large amount of time required to fit datasets of neura
l fields.In this work we propose a JAX-based library that leverages parallelizat
ion to enable fast optimization of large-scale NeF datasets resulting in a signi
ficant speed-up. With this library we perform a comprehensive study that investi
gates the effects of different hyperparameters on fitting NeFs for downstream ta
sks. In particular we explore the use of a shared initialization the effects of
overtraining and the expressiveness of the network architectures used. Our study
 provides valuable insights on how to train NeFs and offers guidance for optimiz
ing their effectiveness in downstream applications. Finally based on the propose
d library and our analysis we propose Neural Field Arena a benchmark consisting
of neural field variants of popular vision datasets including MNIST CIFAR varian
ts of ImageNet and ShapeNetv2. Our library and the Neural Field Arena will be op
en-sourced to introduce standardized benchmarking and promote further research o
n neural fields.

****
Delving into the Trajectory Long-tail Distribution for Muti-object Tracking
Sijia Chen, En Yu, Jinyang Li, Wenbing Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19341-19351
Multiple Object Tracking (MOT) is a critical area within computer vision with a broad spectrum of practical implementations. Current research has primarily focused on the development of tracking algorithms and enhancement of post-processing techniques. Yet there has been a lack of thorough examination concerning the nature of tracking data it self. In this study we pioneer an exploration into the distribution patterns of tracking data and identify a pronounced long-tail distribution issue within existing MOT datasets. We note a significant imbalance in the distribution of trajectory lengths across different pedestrians a phenomenon we refer to as "pedestrians trajectory long-tail distribution". Addressing this challenge we introduce a bespoke strategy designed to mitigate the effects of this skewed distribution. Specifically we propose two data augmentation strategies including Stationary Camera View Data Augmentation (SVA) and Dynamic Camera View Data Augmentation (DVA) designed for viewpoint states and the Group Softmax (GS) module for Re-ID. SVA is to backtrack and predict the pedestrian trajectory of tail classes and DVA is to use diffusion model to change the background of the scene. GS divides the pedestrians into unrelated groups and performs softmax operation on each group individually. Our proposed strategies can be integrated into numerous existing tracking systems and extensive experimentation validates the efficacy of our method in reducing the influence of long-tail distribution on multi-object tracking performance. The code is available at https://github.com/chen-si-jia/Trajectory-Long-tail-Distribution-for-MOT.
****
Tri-Modal Motion Retrieval by Learning a Joint Embedding Space
Kangning Yin, Shihao Zou, Yuxuan Ge, Zheng Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1596-1605
Text-to-motion tasks have been the focus of recent advancements in the human motion domain. However the performance of text-to-motion tasks have not reached its potential primarily due to the lack of motion datasets and the pronounced gap between the text and motion modalities. To mitigate this challenge we introduce VLMA a novel Video-Language-Motion Alignment method. This approach leverages human-centric videos as an intermediary modality effectively bridging the divide between text and motion. By employing contrastive learning we construct a cohesive embedding space across the three modalities. Furthermore we incorporate a motion reconstruction branch ensuring that the resulting motion remains closely aligned with its original trajectory. Experimental evaluations on the HumanML3D and KIT-ML datasets demonstrate the superiority of our method in comparison to existing approaches. Furthermore we introduce a novel task termed video-to-motion retrieval designed to facilitate the seamlessxt eraction of corresponding 3D motions from an RGB video. Supplementary experiments demonstrate that our model is extensible to real-world human-centric videos offering a valuable complement to the pose estimation task.
****
Seg2Reg: Differentiable 2D Segmentation to 1D Regression Rendering for 360 Room Layout Reconstruction
Cheng Sun, Wei-En Tai, Yu-Lin Shih, Kuan-Wei Chen, Yong-Jing Syu, Kent Selwyn The, Yu-Chiang Frank Wang, Hwann-Tzong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10435-10445
State-of-the-art single-view 360 room layout reconstruction methods formulate the problem as a high-level 1D (per-column) regression task. On the other hand traditional low-level 2D layout segmentation is simpler to learn and can represent occluded regions but it requires complex post-processing for the targeting layout polygon and sacrifices accuracy. We present Seg2Reg to render 1D layout depth regression from the 2D segmentation map in a differentiable and occlusion-aware way marrying the merits of both sides. Specifically our model predicts floor-plan density for the input equirectangular 360 image. Formulating the 2D layout representation as a density field enables us to employ 'flattened' volume rendering

to form 1D layout depth regression. In addition we propose a novel 3D warping augmentation on layout to improve generalization. Finally we re-implement recent room layout reconstruction methods into our codebase for benchmarking and explore modern backbones and training techniques to serve as the strong baseline. The code is at https: //PanoLayoutStudio.github.io .
****
Strong Transferable Adversarial Attacks via Ensembled Asymptotically Normal Distribution Learning
Zhengwei Fang, Rui Wang, Tao Huang, Liping Jing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24841-24850
Strong adversarial examples are crucial for evaluating and enhancing the robustness of deep neural networks. However the performance of popular attacks is usually sensitive for instance to minor image transformations stemming from limited information -- typically only one input example a handful of white-box source models and undefined defense strategies. Hence the crafted adversarial examples are prone to overfit the source model which hampers their transferability to unknown architectures. In this paper we propose an approach named Multiple Asymptotically Normal Distribution Attacks (MultiANDA) which explicitly characterize adversarial perturbations from a learned distribution. Specifically we approximate the posterior distribution over the perturbations by taking advantage of the asymptotic normality property of stochastic gradient ascent (SGA) then employ the deep ensemble strategy as an effective proxy for Bayesian marginalization in this process aiming to estimate a mixture of Gaussians that facilitates a more thorough exploration of the potential optimization space. The approximated posterior essentially describes the stationary distribution of SGA iterations which captures the geometric information around the local optimum. Thus MultiANDA allows drawing an unlimited number of adversarial perturbations for each input and reliably maintains the transferability. Our proposed method outperforms ten state-of-the-art black-box attacks on deep learning models with or without defenses through extensive experiments on seven normally trained and seven defense models.
****
Spanning Training Progress: Temporal Dual-Depth Scoring (TDDS) for Enhanced Dataset Pruning
Xin Zhang, Jiawei Du, Yunsong Li, Weiying Xie, Joey Tianyi Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26223-26232
Dataset pruning aims to construct a coreset capable of achieving performance comparable to the original full dataset. Most existing dataset pruning methods rely on snapshot-based criteria to identify representative samples often resulting in poor generalization across various pruning and cross-architecture scenarios. Recent studies have addressed this issue by expanding the scope of training dynamics considered including factors such as forgetting event and probability change typically using an averaging approach. However these works struggle to integrate a broader range of training dynamics without overlooking well-generalized samples which may not be sufficiently highlighted in an averaging manner. In this study we propose a novel dataset pruning method termed as Temporal Dual-Depth Scoring (TDDS) to tackle this problem. TDDS utilizes a dual-depth strategy to achieve a balance between incorporating extensive training dynamics and identifying representative samples for dataset pruning. In the first depth we estimate the series of each sample's individual contributions spanning the training progress ensuring comprehensive integration of training dynamics. In the second depth we focus on the variability of the sample-wise contributions identified in the first depth to highlight well-generalized samples. Extensive experiments conducted on CIFAR and ImageNet datasets verify the superiority of TDDS over previous SOTA methods. Specifically on CIFAR-100 our method achieves 54.51% accuracy with only 10% training data surpassing baselines methods by more than 12.69%. Our codes are available at https://github.com/zhangxin-xd/Dataset-Pruning-TDDS.
****
UniMix: Towards Domain Adaptive and Generalizable LiDAR Semantic Segmentation in Adverse Weather

Haimei Zhao, Jing Zhang, Zhuo Chen, Shanshan Zhao, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14781-14791
LiDAR semantic segmentation (LSS) is a critical task in autonomous driving and has achieved promising progress. However prior LSS methods are conventionally investigated and evaluated on datasets within the same domain in clear weather. The robustness of LSS models in unseen scenes and all weather conditions is crucial for ensuring safety and reliability in real applications. To this end we propose UniMix a universal method that enhances the adaptability and generalizability of LSS models. UniMix first leverages physically valid adverse weather simulation to construct a Bridge Domain which serves to bridge the domain gap between the clear weather scenes and the adverse weather scenes. Then a Universal Mixing operator is defined regarding spatial intensity and semantic distributions to create the intermediate domain with mixed samples from given domains. Integrating the proposed two techniques into a teacher-student framework UniMix efficiently mitigates the domain gap and enables LSS models to learn weather-robust and domain-invariant representations. We devote UniMix to two main setups: 1) unsupervised domain adaption adapting the model from the clear weather source domain to the adverse weather target domain; 2) domain generalization learning a model that generalizes well to unseen scenes in adverse weather. Extensive experiments validate the effectiveness of UniMix across different tasks and datasets all achieving superior performance over state-of-the-art methods. The code will be released.
****
Visual Delta Generator with Large Multi-modal Models for Semi-supervised Composed Image Retrieval
Young Kyun Jang, Donghyun Kim, Zihang Meng, Dat Huynh, Ser-Nam Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16805-16814
Composed Image Retrieval (CIR) is a task that retrieves images similar to a query based on a provided textual modification. Current techniques rely on supervised learning for CIR models using labeled triplets of the <reference image text target image>. These specific triplets are not as commonly available as simple image-text pairs limiting the widespread use of CIR and its scalability. On the other hand zero-shot CIR can be relatively easily trained with image-caption pairs without considering the image-to-image relation but this approach tends to yield lower accuracy. We propose a new semi-supervised CIR approach where we search for a reference and its related target images in auxiliary data and learn our large language model-based Visual Delta Generator (VDG) to generate text describing the visual difference (i.e. visual delta) between the two. VDG equipped with fluent language knowledge and being model agnostic can generate pseudo triplets to boost the performance of CIR models. Our approach significantly improves the existing supervised learning approaches and achieves state-of-the-art results on the CIR benchmarks.
****
Selective Interpretable and Motion Consistent Privacy Attribute Obfuscation for Action Recognition
Filip Ilic, He Zhao, Thomas Pock, Richard P. Wildes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18730-18739
Concerns for the privacy of individuals captured in public imagery have led to privacy-preserving action recognition. Existing approaches often suffer from issues arising through obfuscation being applied globally and a lack of interpretability. Global obfuscation hides privacy sensitive regions but also contextual regions important for action recognition. Lack of interpretability erodes trust in these new technologies. We highlight the limitations of current paradigms and propose a solution: Human selected privacy templates that yield interpretability by design an obfuscation scheme that selectively hides attributes and also induces temporal consistency which is important in action recognition. Our approach is architecture agnostic and directly modifies input imagery while existing approaches generally require architecture training. Our approach offers more flexibili

ty as no retraining is required and outperforms alternatives on three widely use
d datasets.
****
HiPose: Hierarchical Binary Surface Encoding and Correspondence Pruning for RGB-
D 6DoF Object Pose Estimation
Yongliang Lin, Yongzhi Su, Praveen Nathan, Sandeep Inuganti, Yan Di, Martin Sund
ermeyer, Fabian Manhardt, Didier Stricker, Jason Rambach, Yu Zhang; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 10148-10158
In this work we present a novel dense-correspondence method for 6DoF object pose
 estimation from a single RGB-D image. While many existing data-driven methods a
chieve impressive performance they tend to be time-consuming due to their relian
ce on rendering-based refinement approaches. To circumvent this limitation we pr
esent HiPose which establishes 3D-3D correspondences in a coarse-to-fine manner
with a hierarchical binary surface encoding. Unlike previous dense-correspondenc
e methods we estimate the correspondence surface by employing point-to-surface m
atching and iteratively constricting the surface until it becomes a corresponden
ce point while gradually removing outliers. Extensive experiments on public benc
hmarks LM-O YCB-V and T-Less demonstrate that our method surpasses all refinemen
t-free methods and is even on par with expensive refinement-based approaches. Cr
ucially our approach is computationally efficient and enables real-time critical
 applications with high accuracy requirements.
****
DiffForensics: Leveraging Diffusion Prior to Image Forgery Detection and Localiz
ation
Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, Bin Li; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12765
-12774
As manipulating images may lead to misinterpretation of the visual content addre
ssing the image forgery detection and localization (IFDL) problem has drawn seri
ous public concerns. In this work we propose a simple assumption that the effect
ive forensic method should focus on the mesoscopic properties of images. Based o
n the assumption a novel two-stage self-supervised framework leveraging the diff
usion model for IFDL task i.e. DiffForensics is proposed in this paper. The Diff
Forensics begins with self-supervised denoising diffusion paradigm equipped with
 the module of encoder-decoder structure by freezing the pre-trained encoder (e.
g. in ADE-20K) to inherit macroscopic features for general image characteristics
 while encouraging the decoder to learn microscopic feature representation of im
ages enforcing the whole model to focus the mesoscopic representations. The pre-
trained model as a prior is then further fine-tuned for IFDL task with the custo
mized Edge Cue Enhancement Module (ECEM) which progressively highlights the boun
dary features within the manipulated regions thereby refining tampered area loca
lization with better precision. Extensive experiments on several public challeng
ing datasets demonstrate the effectiveness of the proposed method compared with
other state-of-the-art methods. The proposed DiffForensics could significantly i
mprove the model's capabilities for both accurate tamper detection and precise t
amper localization while concurrently elevating its generalization and robustnes
s.
****
CoSeR: Bridging Image and Language for Cognitive Super-Resolution
Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youlian
g Yan, Yujiu Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2024, pp. 25868-25878
Existing super-resolution (SR) models primarily focus on restoring local texture
 details often neglecting the global semantic information within the scene. This
 oversight can lead to the omission of crucial semantic details or the introduct
ion of inaccurate textures during the recovery process. In our work we introduce
 the Cognitive Super-Resolution (CoSeR) framework empowering SR models with the
capacity to comprehend low-resolution images. We achieve this by marrying image
appearance and language understanding to generate a cognitive embedding which no

t only activates prior information from large text-to-image diffusion models but
 also facilitates the generation of high-quality reference images to optimize th
e SR process. To further improve image fidelity we propose a novel condition inj
ection scheme called "All-in-Attention" consolidating all conditional informatio
n into a single module. Consequently our method successfully restores semantical
ly correct and photorealistic details demonstrating state-of-the-art performance
 across multiple benchmarks. Project page: https://coser-main.github.io/
****
Geometry-aware Reconstruction and Fusion-refined Rendering for Generalizable Neu
ral Radiance Fields
Tianqi Liu, Xinyi Ye, Min Shi, Zihao Huang, Zhiyu Pan, Zhan Peng, Zhiguo Cao; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 7654-7663
Generalizable NeRF aims to synthesize novel views for unseen scenes. Common prac
tices involve constructing variance-based cost volumes for geometry reconstructi
on and encoding 3D descriptors for decoding novel views. However existing method
s show limited generalization ability in challenging conditions due to inaccurat
e geometry sub-optimal descriptors and decoding strategies. We address these iss
ues point by point. First we find the variance-based cost volume exhibits failur
e patterns as the features of pixels corresponding to the same point can be inco
nsistent across different views due to occlusions or reflections. We introduce a
n Adaptive Cost Aggregation (ACA) approach to amplify the contribution of consis
tent pixel pairs and suppress inconsistent ones. Unlike previous methods that so
lely fuse 2D features into descriptors our approach introduces a Spatial-View Ag
gregator (SVA) to incorporate 3D context into descriptors through spatial and in
ter-view interaction. When decoding the descriptors we observe the two existing
decoding strategies excel in different areas which are complementary. A Consiste
ncy-Aware Fusion (CAF) strategy is proposed to leverage the advantages of both.
We incorporate the above ACA SVA and CAF into a coarse-to-fine framework termed
Geometry-aware Reconstruction and Fusion-refined Rendering (GeFu). GeFu attains
state-of-the-art performance across multiple datasets.
****
Boosting Self-Supervision for Single-View Scene Completion via Knowledge Distill
ation
Keonhee Han, Dominik Muhle, Felix Wimbauer, Daniel Cremers; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
9837-9847
Inferring scene geometry from images via Structure from Motion is a long-standin
g and fundamental problem in computer vision. While classical approaches and mor
e recently depth map predictions only focus on the visible parts of a scene the
task of scene completion aims to reason about geometry even in occluded regions.
 With the popularity of NeRF implicit representations also became popular for sc
ene completion by predicting so-called density fields. Unlike explicit approache
s e.g. voxel-based methods density fields also allow for accurate depth predicti
on and novel-view synthesis via image-based rendering. In this work we propose t
o fuse the scene reconstruction from multiple images and distill this knowledge
into a more accurate single-view scene reconstruction. To this end we propose MV
BTS to fuse density fields from multiple posed images trained fully self-supervi
sed only from image data. Using knowledge distillation we use MVBTS to train a s
ingle-view scene completion network via direct supervision called KDBTS. It achi
eves state-of-the-art performance on occupancy prediction especially in occluded
 regions.
****
PromptKD: Unsupervised Prompt Distillation for Vision-Language Models
Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, Jian Yang; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 26617-26626
Prompt learning has emerged as a valuable technique in enhancing vision-language
 models (VLMs) such as CLIP for downstream tasks in specific domains. Existing w
ork mainly focuses on designing various learning forms of prompts neglecting the

potential of prompts as effective distillers for learning from larger teacher m odels. In this paper we introduce an unsupervised domain prompt distillation fra mework which aims to transfer the knowledge of a larger teacher model to a light weight target model through prompt-driven imitation using unlabeled domain image s. Specifically our framework consists of two distinct stages. In the initial st age we pre-train a large CLIP teacher model using domain (few-shot) labels. Afte r pre-training we leverage the unique decoupled-modality characteristics of CLIP by pre-computing and storing the text features as class vectors only once throu gh the teacher text encoder. In the subsequent stage the stored class vectors ar e shared across teacher and student image encoders for calculating the predicted logits. Further we align the logits of both the teacher and student models via KL divergence encouraging the student image encoder to generate similar probabil ity distributions to the teacher through the learnable prompts. The proposed pro mpt distillation process eliminates the reliance on labeled data enabling the al gorithm to leverage a vast amount of unlabeled images within the domain. Finally the well-trained student image encoders and pre-stored text features (class vec tors) are utilized for inference. To our best knowledge we are the first to (1) perform unsupervised domain-specific prompt-driven knowledge distillation for CL IP and (2) establish a practical pre-storing mechanism of text features as share d class vectors between teacher and student. Extensive experiments on 11 dataset s demonstrate the effectiveness of our method. Code is publicly available at htt ps://github.com/zhengli97/PromptKD.
****
VideoBooth: Diffusion-based Video Generation with Image Prompts
Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Cha nge Loy, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2024, pp. 6689-6700
Text-driven video generation witnesses rapid progress. However merely using text prompts is not enough to depict the desired subject appearance that accurately aligns with users' intents especially for customized content creation. In this p aper we study the task of video generation with image prompts which provide more accurate and direct content control beyond the text prompts. Specifically we pr opose a feed-forward framework VideoBooth with two dedicated designs: 1) We prop ose to embed image prompts in a coarse-to-fine manner. Coarse visual embeddings from image encoder provide high-level encodings of image prompts while fine visu al embeddings from the proposed attention injection module provide multi-scale a nd detailed encoding of image prompts. These two complementary embeddings can fa ithfully capture the desired appearance. 2) In the attention injection module at fine level multi-scale image prompts are fed into different cross-frame attenti on layers as additional keys and values. This extra spatial information refines the details in the first frame and then it is propagated to the remaining frames which maintains temporal consistency. Extensive experiments demonstrate that Vi deoBooth achieves state-of-the-art performance in generating customized high-qua lity videos with subjects specified in image prompts. Notably VideoBooth is a ge neralizable framework where a single model works for a wide range of image promp ts with only feed-forward passes.
****
Robust Overfitting Does Matter: Test-Time Adversarial Purification With FGSM
Linyu Tang, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24347-24356
Numerous studies have demonstrated the susceptibility of deep neural networks (D NNs) to subtle adversarial perturbations prompting the development of many advan ced adversarial defense methods aimed at mitigating adversarial attacks. Current defense strategies usually train DNNs for a specific adversarial attack method and can achieve good robustness in defense against this type of adversarial atta ck. Nevertheless when subjected to evaluations involving unfamiliar attack modal ities empirical evidence reveals a pronounced deterioration in the robustness of DNNs. Meanwhile there is a trade-off between the classification accuracy of cle an examples and adversarial examples. Most defense methods often sacrifice the a ccuracy of clean examples in order to improve the adversarial robustness of DNNs

. To alleviate these problems and enhance the overall robust generalization of DNNs we propose the Test-Time Pixel-Level Adversarial Purification (TPAP) method. This approach is based on the robust overfitting characteristic of DNNs to the fast gradient sign method (FGSM) on training and test datasets. It utilizes FGSM for adversarial purification to process images for purifying unknown adversarial perturbations from pixels at testing time in a "counter changes with changelessness" manner thereby enhancing the defense capability of DNNs against various unknown adversarial attacks. Extensive experimental results show that our method can effectively improve both overall robust generalization of DNNs notably over previous methods. Code is available https://github.com/tly18/TPAP.
****
Sparse Global Matching for Video Frame Interpolation with Large Motion
Chunxu Liu, Guozhen Zhang, Rui Zhao, Limin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19125-19134
Large motion poses a critical challenge in Video Frame Interpolation (VFI) task. Existing methods are often constrained by limited receptive fields resulting in sub-optimal performance when handling scenarios with large motion. In this paper we introduce a new pipeline for VFI which can effectively integrate global-level information to alleviate issues associated with large motion. Specifically we first estimate a pair of initial intermediate flows using a high-resolution feature map for extracting local details. Then we incorporate a sparse global matching branch to compensate for flow estimation which consists of identifying flaws in initial flows and generating sparse flow compensation with a global receptive field. Finally we adaptively merge the initial flow estimation with global flow compensation yielding a more accurate intermediate flow. To evaluate the effectiveness of our method in handling large motion we carefully curate a more challenging subset from commonly used benchmarks. Our method demonstrates the state-of-the-art performance on these VFI subsets with large motion.
****
ExtDM: Distribution Extrapolation Diffusion Model for Video Prediction
Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, Jufeng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19310-19320
Video prediction is a challenging task due to its nature of uncertainty especially for forecasting a long period. To model the temporal dynamics advanced methods benefit from the recent success of diffusion models and repeatedly refine the predicted future frames with 3D spatiotemporal U-Net. However there exists a gap between the present and future and the repeated usage of U-Net brings a heavy computation burden. To address this we propose a diffusion-based video prediction method that predicts future frames by extrapolating the present distribution of features namely ExtDM. Specifically our method consists of three components: (i) a motion autoencoder conducts a bijection transformation between video frames and motion cues; (ii) a layered distribution adaptor module extrapolates the present features in the guidance of Gaussian distribution; (iii) a 3D U-Net architecture specialized for jointly fusing guidance and features among the temporal dimension by spatiotemporal-window attention. Extensive experiments on five popular benchmarks covering short- and long-term video prediction verify the effectiveness of ExtDM.
****
Modality-Collaborative Test-Time Adaptation for Action Recognition
Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26732-26741
Video-based Unsupervised Domain Adaptation (VUDA) method improves the generalization of the video model enabling it to be applied to action recognition tasks in different environments. However these methods require continuous access to source data during the adaptation process which are impractical in real scenarios where the source videos are not available with concerns in transmission efficiency or privacy issues. To address this problem in this paper we propose to solve the Multimodal Video Test-Time Adaptation task (MVTTA). Existing image-based TTA m

ethods cannot be directly applied to this task because video have domain shift i
n multimodal and temporal which brings difficulties to adaptation. To address th
e above challenges we propose a Modality-Collaborative Test-Time Adaptation (MC-
TTA) Network. We maintain teacher and student memory banks respectively for gene
rating pseudo-prototypes and target-prototypes. In the teacher model we propose
Self-assembled Source-friendly Feature Reconstruction (SSFR) module to encourage
 the teacher memory bank to store features that are more likely to be consistent
 with the source distribution. Through multimodal prototype alignment and cross-
modal relative consistency our method can effectively alleviate domain shift in
videos. We evaluate the proposed model on four public video datasets. The result
s show that our model outperforms existing state-of-the-art methods.
****
SCULPT: Shape-Conditioned Unpaired Learning of Pose-dependent Clothed and Textur
ed Human Meshes
Soubhik Sanyal, Partha Ghosh, Jinlong Yang, Michael J. Black, Justus Thies, Timo
 Bolkart; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2024, pp. 2362-2371
We present SCULPT a novel 3D generative model for clothed and textured 3D meshes
 of humans. Specifically we devise a deep neural network that learns to represen
t the geometry and appearance distribution of clothed human bodies. Training suc
h a model is challenging as datasets of textured 3D meshes for humans are limite
d in size and accessibility. Our key observation is that there exist medium-size
d 3D scan datasets like CAPE as well as large-scale 2D image datasets of clothed
 humans and multiple appearances can be mapped to a single geometry. To effectiv
ely learn from the two data modalities we propose an unpaired learning procedure
 for pose-dependent clothed and textured human meshes. Specifically we learn a p
ose-dependent geometry space from 3D scan data. We represent this as per vertex
displacements w.r.t. the SMPL model. Next we train a geometry conditioned textur
e generator in an unsupervised way using the 2D image data. We use intermediate
activations of the learned geometry model to condition our texture generator. To
 alleviate entanglement between pose and clothing type and pose and clothing app
earance we condition both the texture and geometry generators with attribute lab
els such as clothing types for the geometry and clothing colors for the texture
generator. We automatically generated these conditioning labels for the 2D image
s based on the visual question-answering model BLIP and CLIP. We validate our me
thod on the SCULPT dataset and compare to state-of-the-art 3D generative models
for clothed human bodies. Our code and data can be found at https://sculpt.is.tu
e.mpg.de.
****
Point Segment and Count: A Generalized Framework for Object Counting
Zhizhong Huang, Mingliang Dai, Yi Zhang, Junping Zhang, Hongming Shan; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2024, pp. 17067-17076
Class-agnostic object counting aims to count all objects in an image with respec
t to example boxes or class names a.k.a few-shot and zero-shot counting. In this
 paper we propose a generalized framework for both few-shot and zero-shot object
 counting based on detection. Our framework combines the superior advantages of
two foundation models without compromising their zero-shot capability: (i) SAM t
o segment all possible objects as mask proposals and (ii) CLIP to classify propo
sals to obtain accurate object counts. However this strategy meets the obstacles
 of efficiency overhead and the small crowded objects that cannot be localized a
nd distinguished. To address these issues our framework termed PseCo follows thr
ee steps: point segment and count. Specifically we first propose a class-agnosti
c object localization to provide accurate but least point prompts for SAM which
consequently not only reduces computation costs but also avoids missing small ob
jects. Furthermore we propose a generalized object classification that leverages
 CLIP image/text embeddings as the classifier following a hierarchical knowledge
 distillation to obtain discriminative classifications among hierarchical mask p
roposals. Extensive experimental results on FSC-147 COCO and LVIS demonstrate th
at PseCo achieves state-of-the-art performance in both few-shot/zero-shot object

counting/detection.
****
Small Steps and Level Sets: Fitting Neural Surface Models with Point Guidance
Chamin Hewa Koneputugodage, Yizhak Ben-Shabat, Dylan Campbell, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21456-21465
A neural signed distance function (SDF) is a convenient shape representation for many tasks such as surface reconstruction editing and generation. However neural SDFs are difficult to fit to raw point clouds such as those sampled from the surface of a shape by a scanner. A major issue occurs when the shape's geometry is very different from the structural biases implicit in the network's initialization. In this case we observe that the standard loss formulation does not guide the network towards the correct SDF values. We circumvent this problem by introducing guiding points and use them to steer the optimization towards the true shape via small incremental changes for which the loss formulation has a good descent direction. We show that this point-guided homotopy-based optimization scheme facilitates a deformation from an easy problem to the difficult reconstruction problem. We also propose a metric to quantify the difference in surface geometry between a target shape and an initial surface which helps indicate whether the standard loss formulation is guiding towards the target shape. Our method outperforms previous state-of-the-art approaches with large improvements on shapes identified by this metric as particularly challenging.
****
Domain-Agnostic Mutual Prompting for Unsupervised Domain Adaptation
Zhekai Du, Xinyao Li, Fengling Li, Ke Lu, Lei Zhu, Jingjing Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23375-23384
Conventional Unsupervised Domain Adaptation (UDA) strives to minimize distribution discrepancy between domains which neglects to harness rich semantics from data and struggles to handle complex domain shifts. A promising technique is to leverage the knowledge of large-scale pre-trained vision-language models for more guided adaptation. Despite some endeavors current methods often learn textual prompts to embed domain semantics for source and target domains separately and perform classification within each domain limiting cross-domain knowledge transfer. Moreover prompting only the language branch lacks flexibility to adapt both modalities dynamically. To bridge this gap we propose Domain-Agnostic Mutual Prompting (DAMP) to exploit domain-invariant semantics by mutually aligning visual and textual embeddings. Specifically the image contextual information is utilized to prompt the language branch in a domain-agnostic and instance-conditioned way. Meanwhile visual prompts are imposed based on the domain-agnostic textual prompt to elicit domain-invariant visual embeddings. These two branches of prompts are learned mutually with a cross-attention module and regularized with a semantic-consistency loss and an instance-discrimination contrastive loss. Experiments on three UDA benchmarks demonstrate the superiority of DAMP over state-of-the-art approaches.
****
PTT: Point-Trajectory Transformer for Efficient Temporal 3D Object Detection
Kuan-Chih Huang, Weijie Lyu, Ming-Hsuan Yang, Yi-Hsuan Tsai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14938-14947
Recent temporal LiDAR-based 3D object detectors achieve promising performance based on the two-stage proposal-based approach. They generate 3D box candidates from the first-stage dense detector followed by different temporal aggregation methods. However these approaches require per-frame objects or whole point clouds posing challenges related to memory bank utilization. Moreover point clouds and trajectory features are combined solely based on concatenation which may neglect effective interactions between them. In this paper we propose a point-trajectory transformer with long short-term memory for efficient temporal 3D object detection. To this end we only utilize point clouds of current-frame objects and their historical trajectories as input to minimize the memory bank storage requiremen

t. Furthermore we introduce modules to encode trajectory features focusing on long short-term and future-aware perspectives and then effectively aggregate them with point cloud features. We conduct extensive experiments on the large-scale Waymo dataset to demonstrate that our approach performs well against state-of-the-art methods. The source codes and trained models will be made publicly available. Code and models will be made publicly available at https://github.com/kuanchihhuang/PTT.
****
Generative Proxemics: A Prior for 3D Social Interaction from Images
Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, Angjoo Kanazawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9687-9697
Social interaction is a fundamental aspect of human behavior and communication. The way individuals position themselves in relation to others also known as proxemics conveys social cues and affects the dynamics of social interaction. Reconstructing such interaction from images presents challenges because of mutual occlusion and the limited availability of large training datasets. To address this we present a novel approach that learns a prior over the 3D proxemics two people in close social interaction and demonstrate its use for single-view 3D reconstruction. We start by creating 3D training data of interacting people using image datasets with contact annotations. We then model the proxemics using a novel denoising diffusion model called BUDDI that learns the joint distribution over the poses of two people in close social interaction. Sampling from our generative proxemics model produces realistic 3D human interactions which we validate through a perceptual study. We use BUDDI in reconstructing two people in close proximity from an image without any contact annotation via an optimization approach that uses the diffusion model as a prior. Our approach recovers accurate 3D social interactions from noisy initial estimates outperforming state-of-the-art methods. Our code data and model are available at: muelea.github.io/buddi.
****
A Simple and Effective Point-based Network for Event Camera 6-DOFs Pose Relocalization
Hongwei Ren, Jiadong Zhu, Yue Zhou, Haotian Fu, Yulong Huang, Bojun Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18112-18121
Event cameras exhibit remarkable attributes such as high dynamic range asynchronicity and low latency making them highly suitable for vision tasks that involve high-speed motion in challenging lighting conditions. These cameras implicitly capture movement and depth information in events making them appealing sensors for Camera Pose Relocalization (CPR) tasks. Nevertheless existing CPR networks based on events neglect the pivotal fine-grained temporal information in events resulting in unsatisfactory performance. Moreover the energy-efficient features are further compromised by the use of excessively complex models hindering efficient deployment on edge devices. In this paper we introduce PEPNet a simple and effective point-based network designed to regress six degrees of freedom (6-DOFs) event camera poses. We rethink the relationship between the event camera and CPR tasks leveraging the raw Point Cloud directly as network input to harness the high-temporal resolution and inherent sparsity of events. PEPNet is adept at abstracting the spatial and implicit temporal features through hierarchical structure and explicit temporal features by Attentive Bi-directional Long Short-Term Memory (A-Bi-LSTM). By employing a carefully crafted lightweight design PEPNet delivers state-of-the-art (SOTA) performance on both indoor and outdoor datasets with meager computational resources. Specifically PEPNet attains a significant 38% and 33% performance improvement on the random split IJRR and M3ED datasets respectively. Moreover the lightweight design version PEPNet_ tiny accomplishes results comparable to the SOTA while employing a mere 0.5% of the parameters.
****
Semantic-Aware Multi-Label Adversarial Attacks
Hassan Mahmood, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24251-24262

Despite its importance generating attacks for multi label learning (MLL) models has received much less attention compared to multi-class recognition. Attacking an MLL model by optimizing a loss on the target set of labels has often the undesired consequence of changing the predictions for other labels. On the other hand adding a loss on the remaining labels to keep them fixed leads to highly negatively correlated gradient directions reducing the attack effectiveness. In this paper we develop a framework for crafting effective and semantic aware adversarial attacks for MLL. First to obtain an attack that leads to semantically consistent predictions across all labels we find a minimal superset of the target labels referred to as consistent target set. To do so we develop an efficient search algorithm over a knowledge graph which encodes label dependencies. Next we propose an optimization that searches for an attack that modifies the predictions of labels in the consistent target set while ensuring other labels will not get affected. This leads to an efficient algorithm that projects the gradient of the consistent target set loss onto the orthogonal direction of the gradient of the loss on other labels. Our framework can generate attacks on different target set sizes and for MLL with thousands of labels (as in OpenImages). Finally by extensive experiments on three datasets and several MLL models we show that our method generates both successful and semantically consistent attacks.
****

EasyDrag: Efficient Point-based Manipulation on Diffusion Models
Xingzhong Hou, Boxiao Liu, Yi Zhang, Jihao Liu, Yu Liu, Haihang You; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8404-8413
Generative models are gaining increasing popularity and the demand for precisely generating images is on the rise. However generating an image that perfectly aligns with users' expectations is extremely challenging. The shapes of objects the poses of animals the structures of landscapes and more may not match the user's desires and this applies to real images as well. This is where point-based image editing becomes essential. An excellent image editing method needs to meet the following criteria: user-friendly interaction high performance and good generalization capability. Due to the limitations of StyleGAN DragGAN exhibits limited robustness across diverse scenarios while DragDiffusion lacks user-friendliness due to the necessity of LoRA fine-tuning and masks. In this paper we introduce a novel interactive point-based image editing framework called EasyDrag that leverages pretrained diffusion models to achieve high-quality editing outcomes and user-friendship. Extensive experimentation demonstrates that our approach surpasses DragDiffusion in terms of both image quality and editing precision for point-based image manipulation tasks.
****

Region-Based Representations Revisited
Michal Shlapentokh-Rothman, Ansel Blume, Yao Xiao, Yuqun Wu, Sethuraman TV, Heyi Tao, Jae Yong Lee, Wilfredo Torres, Yu-Xiong Wang, Derek Hoiem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17107-17116
We investigate whether region-based representations are effective for recognition. Regions were once a mainstay in recognition approaches but pixel and patch-based features are now used almost exclusively. We show that recent class-agnostic segmenters like SAM can be effectively combined with strong unsupervised representations like DINOv2 and used for a wide variety of tasks including semantic segmentation object-based image retrieval and multi-image analysis. Once the masks and features are extracted these representations even with linear decoders enable competitive performance making them well suited to applications that require custom queries. The compactness of the representation also makes it well-suited to video analysis and other problems requiring inference across many images.
****

GenH2R: Learning Generalizable Human-to-Robot Handover via Scalable Simulation Demonstration and Imitation
Zifan Wang, Junyu Chen, Ziqing Chen, Pengwei Xie, Rui Chen, Li Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202

4, pp. 16362-16372
This paper presents GenH2R a framework for learning generalizable vision-based h
uman-to-robot (H2R) handover skills. The goal is to equip robots with the abilit
y to reliably receive objects with unseen geometry handed over by humans in vari
ous complex trajectories. We acquire such generalizability by learning H2R hando
ver at scale with a comprehensive solution including procedural simulation asset
s creation automated demonstration generation and effective imitation learning.
We leverage large-scale 3D model repositories dexterous grasp generation methods
 and curve-based 3D animation to create an H2R handover simulation environment n
amed GenH2R-Sim surpassing the number of scenes in existing simulators by three
orders of magnitude. We further introduce a distillation-friendly demonstration
generation method that automatically generates a million high-quality demonstrat
ions suitable for learning. Finally we present a 4D imitation learning method au
gmented by a future forecasting objective to distill demonstrations into a visuo
-motor handover policy. Experimental evaluations in both simulators and the real
 world demonstrate significant improvements (at least +10% success rate) over ba
selines in all cases.
****
Modality-Agnostic Structural Image Representation Learning for Deformable Multi-
Modality Medical Image Registration
Tony C. W. Mok, Zi Li, Yunhao Bai, Jianpeng Zhang, Wei Liu, Yan-Jie Zhou, Ke Yan
, Dakai Jin, Yu Shi, Xiaoli Yin, Le Lu, Ling Zhang; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11215-11
225
Establishing dense anatomical correspondence across distinct imaging modalities
is a foundational yet challenging procedure for numerous medical image analysis
studies and image-guided radiotherapy. Existing multi-modality image registratio
n algorithms rely on statistical-based similarity measures or local structural i
mage representations. However the former is sensitive to locally varying noise w
hile the latter is not discriminative enough to cope with complex anatomical str
uctures in multimodal scans causing ambiguity in determining the anatomical corr
espondence across scans with different modalities. In this paper we propose a mo
dality-agnostic structural representation learning method which leverages Deep N
eighbourhood Self-similarity (DNS) and anatomy-aware contrastive learning to lea
rn discriminative and contrast-invariance deep structural image representations
(DSIR) without the need for anatomical delineations or pre-aligned training imag
es. We evaluate our method on multiphase CT abdomen MR-CT and brain MR T1w-T2w r
egistration. Comprehensive results demonstrate that our method is superior to th
e conventional local structural representation and statistical-based similarity
measures in terms of discriminability and accuracy.
****
Any-Shift Prompting for Generalization over Distributions
Zehao Xiao, Jiayi Shen, Mohammad Mahdi Derakhshani, Shengcai Liao, Cees G. M. Sn
oek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2024, pp. 13849-13860
Image-language models with prompt learning have shown remarkable advances in num
erous downstream vision tasks. Nevertheless conventional prompt learning methods
 overfit the training distribution and lose the generalization ability on the te
st distributions. To improve the generalization across various distribution shif
ts we propose any-shift prompting: a general probabilistic inference framework t
hat considers the relationship between training and test distributions during pr
ompt learning. We explicitly connect training and test distributions in the late
nt space by constructing training and test prompts in a hierarchical architectur
e. Within this framework the test prompt exploits the distribution relationships
 to guide the generalization of the CLIP image-language model from training to a
ny test distribution. To effectively encode the distribution information and the
ir relationships we further introduce a transformer inference network with a pse
udo-shift training mechanism. The network generates the tailored test prompt wit
h both training and test information in a feedforward pass avoiding extra traini
ng costs at test time. Extensive experiments on twenty-three datasets demonstrat

e the effectiveness of any-shift prompting on the generalization over various di
stribution shifts.
****
InterHandGen: Two-Hand Interaction Generation via Cascaded Reverse Diffusion
Jihyun Lee, Shunsuke Saito, Giljoo Nam, Minhyuk Sung, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 527-537
We present InterHandGen a novel framework that learns the generative prior of two-hand interaction. Sampling from our model yields plausible and diverse two-hand shapes in close interaction with or without an object. Our prior can be incorporated into any optimization or learning methods to reduce ambiguity in an ill-posed setup. Our key observation is that directly modeling the joint distribution of multiple instances imposes high learning complexity due to its combinatorial nature. Thus we propose to decompose the modeling of joint distribution into the modeling of factored unconditional and conditional single instance distribution. In particular we introduce a diffusion model that learns the single-hand distribution unconditional and conditional to another hand via conditioning dropout. For sampling we combine anti-penetration and classifier-free guidance to enable plausible generation. Furthermore we establish the rigorous evaluation protocol of two-hand synthesis where our method significantly outperforms baseline generative models in terms of plausibility and diversity. We also demonstrate that our diffusion prior can boost the performance of two-hand reconstruction from monocular in-the-wild images achieving new state-of-the-art accuracy.
****
CPR-Coach: Recognizing Composite Error Actions based on Single-class Training
Shunli Wang, Shuaibing Wang, Dingkang Yang, Mingcheng Li, Haopeng Kuang, Xiao Zhao, Liuzhen Su, Peng Zhai, Lihua Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18782-18792
Fine-grained medical action analysis plays a vital role in improving medical skill training efficiency but it faces the problems of data and algorithm shortage. Cardiopulmonary Resuscitation (CPR) is an essential skill in emergency treatment. Currently the assessment of CPR skills mainly depends on dummies and trainers leading to high training costs and low efficiency. For the first time this paper constructs a vision-based system to complete error action recognition and skill assessment in CPR. Specifically we define 13 types of single-error actions and 74 types of composite error actions during external cardiac compression and then develop a video dataset named CPR-Coach. By taking the CPR-Coach as a benchmark this paper investigates and compares the performance of existing action recognition models based on different data modalities. To solve the unavoidable "Single-class Training & Multi-class Testing" problem we propose a human-cognition-inspired framework named ImagineNet to improve the model's multi-error recognition performance under restricted supervision. Extensive comparison and actual deployment experiments verify the effectiveness of the framework. We hope this work could bring new inspiration to the computer vision and medical skills training communities simultaneously. The dataset and the code are publicly available on https://github.com/Shunli-Wang/CPR-Coach.
****
Video2Game: Real-time Interactive Realistic and Browser-Compatible Environment from a Single Video
Hongchi Xia, Zhi-Hao Lin, Wei-Chiu Ma, Shenlong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4578-4588
Creating high-quality and interactive virtual environments such as games and simulators often involves complex and costly manual modeling processes. In this paper we present Video2Game a novel approach that automatically converts videos of real-world scenes into realistic and interactive game environments. At the heart of our system are three core components: (i) a neural radiance fields (NeRF) module that effectively captures the geometry and visual appearance of the scene; (ii) a mesh module that distills the knowledge from NeRF for faster rendering; and (iii) a physics module that models the interactions and physical dynamics amo

ng the objects. By following the carefully designed pipeline one can construct an interactable and actionable digital replica of the real world. We benchmark our system on both indoor and large-scale outdoor scenes. We show that we can not only produce highly-realistic renderings in real-time but also build interactive games on top.
****

Tackling the Singularities at the Endpoints of Time Intervals in Diffusion Models

Pengze Zhang, Hubery Yin, Chen Li, Xiaohua Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6945-6954

Most diffusion models assume that the reverse process adheres to a Gaussian distribution. However this approximation has not been rigorously validated especially at singularities where t=0 and t=1. Improperly dealing with such singularities leads to an average brightness issue in applications and limits the generation of images with extreme brightness or darkness. We primarily focus on tackling singularities from both theoretical and practical perspectives. Initially we establish the error bounds for the reverse process approximation and showcase its Gaussian characteristics at singularity time steps. Based on this theoretical insight we confirm the singularity at t=1 is conditionally removable while it at t=0 is an inherent property. Upon these significant conclusions we propose a novel plug-and-play method SingDiffusion to address the initial singular time step sampling which not only effectively resolves the average brightness issue for a wide range of diffusion models without extra training efforts but also enhances their generation capability in achieving notable lower FID scores.
****

MatSynth: A Modern PBR Materials Dataset

Giuseppe Vecchio, Valentin Deschaintre; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22109-22118

We introduce MatSynth a dataset of 4000+ CC0 ultra-high resolution PBR materials. Materials are crucial components of virtual relightable assets defining the interaction of light at the surface of geometries. Given their importance significant research effort was dedicated to their representation creation and acquisition. However in the past 6 years most research in material acquisition or generation relied either on the same unique dataset or on company-owned huge library of procedural materials. With this dataset we propose a significantly larger more diverse and higher resolution set of materials than previously publicly available. We carefully discuss the data collection process and demonstrate the benefits of this dataset for material acquisition and generation applications. The complete data further contains metadata with each material's origin license category tags creation method and when available descriptions and physical size as well as 3M+ renderings of the augmented materials in 1K under various environment lightings. The MatSynth dataset is released through the project page at: https://www.gvecchio.com/matsynth.
****

CHAIN: Enhancing Generalization in Data-Efficient GANs via lipsCHitz continuity constrAIned Normalization

Yao Ni, Piotr Koniusz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6763-6774

Generative Adversarial Networks (GANs) significantly advanced image generation but their performance heavily depends on abundant training data. In scenarios with limited data GANs often struggle with discriminator overfitting and unstable training. Batch Normalization (BN) despite being known for enhancing generalization and training stability has rarely been used in the discriminator of Data-Efficient GANs. Our work addresses this gap by identifying a critical flaw in BN: the tendency for gradient explosion during the centering and scaling steps. To tackle this issue we present CHAIN (lipsCHitz continuity constrAIned Normalization) which replaces the conventional centering step with zero-mean regularization and integrates a Lipschitz continuity constraint in the scaling step. CHAIN further enhances GAN training by adaptively interpolating the normalized and unnormalized features effectively avoiding discriminator overfitting. Our theoretical ana

lyses firmly establishes CHAIN's effectiveness in reducing gradients in latent features and weights improving stability and generalization in GAN training. Empirical evidence supports our theory. CHAIN achieves state-of-the-art results in data-limited scenarios on CIFAR-10/100 ImageNet five low-shot and seven high-resolution few-shot image datasets.
****
RTracker: Recoverable Tracking via PN Tree Structured Memory
Yuqing Huang, Xin Li, Zikun Zhou, Yaowei Wang, Zhenyu He, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19038-19047
Existing tracking methods mainly focus on learning better target representation or developing more robust prediction models to improve tracking performance. While tracking performance has significantly improved the target loss issue occurs frequently due to tracking failures complete occlusion or out-of-view situations. However considerably less attention is paid to the self-recovery issue of tracking methods which is crucial for practical applications. To this end we propose a recoverable tracking framework \ourmethod that uses a tree-structured memory to dynamically associate a tracker and a detector to enable self-recovery ability. Specifically we propose a Positive-Negative Tree-structured memory to chronologically store and maintain positive and negative target samples. Upon the PN tree memory we develop corresponding walking rules for determining the state of the target and define a set of control flows to unite the tracker and the detector in different tracking scenarios. Our core idea is to use the support samples of positive and negative target categories to establish a relative distance-based criterion for a reliable assessment of target loss. The favorable performance in comparison against the state-of-the-art methods on numerous challenging benchmarks demonstrates the effectiveness of the proposed algorithm. All the source code and trained models will be released at https://github.com/NorahGreen/RTracker.
****
High-Quality Facial Geometry and Appearance Capture at Home
Yuxuan Han, Junfeng Lyu, Feng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 697-707
Facial geometry and appearance capture have demonstrated tremendous success in 3D scanning real humans in studios. Recent works propose to democratize this technique while keeping the results high quality. However they are still inconvenient for daily usage. In addition they focus on an easier problem of only capturing facial skin. This paper proposes a novel method for high-quality face capture featuring an easy-to-use system and the capability to model the complete face with skin mouth interior hair and eyes. We reconstruct facial geometry and appearance from a single co-located smartphone flashlight sequence captured in a dim room where the flashlight is the dominant light source (e.g. rooms with curtains or at night). To model the complete face we propose a novel hybrid representation to effectively model both eyes and other facial regions along with novel techniques to learn it from images. We apply a combined lighting model to compactly represent real illuminations and exploit a morphable face albedo model as a reflectance prior to disentangle diffuse and specular. Experiments show that our method can capture high-quality 3D relightable scans. Our code will be released.
****
DualAD: Disentangling the Dynamic and Static World for End-to-End Driving
Simon Doll, Niklas Hanselmann, Lukas Schneider, Richard Schulz, Marius Cordts, Markus Enzweiler, Hendrik P. A. Lensch; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14728-14737
State-of-the-art approaches for autonomous driving integrate multiple sub-tasks of the overall driving task into a single pipeline that can be trained in an end-to-end fashion by passing latent representations between the different modules. In contrast to previous approaches that rely on a unified grid to represent the belief state of the scene we propose dedicated representations to disentangle dynamic agents and static scene elements. This allows us to explicitly compensate for the effect of both ego and object motion between consecutive time steps and to flexibly propagate the belief state through time. Furthermore dynamic object

s can not only attend to the input camera images but also directly benefit from the inferred static scene structure via a novel dynamic-static cross-attention. Extensive experiments on the challenging nuScenes benchmark demonstrate the benefits of the proposed dual-stream design especially for modelling highly dynamic agents in the scene and highlight the improved temporal consistency of our approach. Our method titled DualAD not only outperforms independently trained single-task networks but also improves over previous state-of-the-art end-to-end models by a large margin on all tasks along the functional chain of driving.
****
OTE: Exploring Accurate Scene Text Recognition Using One Token
Jianjun Xu, Yuxin Wang, Hongtao Xie, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28327-28336
In this paper we propose a novel framework to fully exploit the potential of a single vector for scene text recognition (STR). Different from previous sequence-to-sequence methods that rely on a sequence of visual tokens to represent scene text images we prove that just one token is enough to characterize the entire text image and achieve accurate text recognition. Based on this insight we introduce a new paradigm for STR called One Token rEcognizer (OTE). Specifically we implement an image-to-vector encoder to extract the fine-grained global semantics eliminating the need for sequential features. Furthermore an elegant yet potent vector-to-sequence decoder is designed to adaptively diffuse global semantics to corresponding character locations enabling both autoregressive and non-autoregressive decoding schemes. By executing decoding within a high-level representational space our vector-to-sequence (V2S) approach avoids the alignment issues between visual tokens and character embeddings prevalent in traditional sequence-to-sequence methods. Remarkably due to introducing character-wise fine-grained information such global tokens also boost the performance of scene text retrieval tasks. Extensive experiments on synthetic and real datasets demonstrate the effectiveness of our method by achieving new state-of-the-art results on various public STR benchmarks. Our code is available at https://github.com/Xu-Jianjun/OTE.
****
MULDE: Multiscale Log-Density Estimation via Denoising Score Matching for Video Anomaly Detection
Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, Mateusz Kozinski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18868-18877
We propose a novel approach to video anomaly detection: we treat feature vectors extracted from videos as realizations of a random variable with a fixed distribution and model this distribution with a neural network. This lets us estimate the likelihood of test videos and detect video anomalies by thresholding the likelihood estimates. We train our video anomaly detector using a modification of denoising score matching a method that injects training data with noise to facilitate modeling its distribution. To eliminate hyperparameter selection we model the distribution of noisy video features across a range of noise levels and introduce a regularizer that tends to align the models for different levels of noise. At test time we combine anomaly indications at multiple noise scales with a Gaussian mixture model. Running our video anomaly detector induces minimal delays as inference requires merely extracting the features and forward-propagating them through a shallow neural network and a Gaussian mixture model. Our experiments on five popular video anomaly detection benchmarks demonstrate state-of-the-art performance both in the object-centric and in the frame-centric setup.
****
Your Image is My Video: Reshaping the Receptive Field via Image-To-Video Differentiable AutoAugmentation and Fusion
Sofia Casarin, Cynthia I. Ugwu, Sergio Escalera, Oswald Lanz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5829-5839
The landscape of deep learning research is moving towards innovative strategies to harness the true potential of data. Traditionally emphasis has been on scalin

g model architectures resulting in large and complex neural networks which can b
e difficult to train with limited computational resources. However independently
 of the model size data quality (i.e. amount and variability) is still a major f
actor that affects model generalization. In this work we propose a novel techniq
ue to exploit available data through the use of automatic data augmentation for
the tasks of image classification and semantic segmentation. We introduce the fi
rst Differentiable Augmentation Search method (DAS) to generate variations of im
ages that can be processed as videos. Compared to previous approaches DAS is ext
remely fast and flexible allowing the search on very large search spaces in less
 than a GPU day. Our intuition is that the increased receptive field in the temp
oral dimension provided by DAS could lead to benefits also to the spatial recept
ive field. More specifically we leverage DAS to guide the reshaping of the spati
al receptive field by selecting task-dependant transformations. As a result comp
ared to standard augmentation alternatives we improve in terms of accuracy on Im
ageNet Cifar10 Cifar100 Tiny-ImageNet Pascal-VOC-2012 and CityScapes datasets wh
en plugging-in our DAS over different light-weight video backbones.
****
PTQ4SAM: Post-Training Quantization for Segment Anything
Chengtao Lv, Hong Chen, Jinyang Guo, Yifu Ding, Xianglong Liu; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 15941-15951
Segment Anything Model (SAM) has achieved impressive performance in many compute
r vision tasks. However as a large-scale model the immense memory and computatio
n costs hinder its practical deployment. In this paper we propose a post-trainin
g quantization (PTQ) framework for Segment Anything Model namely PTQ4SAM. First
we investigate the inherent bottleneck of SAM quantization attributed to the bim
odal distribution in \cls post-Key-Linear activations. We analyze its character
istics from both per-tensor and per-channel perspectives and propose a Bimodal I
ntegration strategy which utilizes a mathematically equivalent sign operation to
 transform the bimodal distribution into a relatively easy-quantized normal dist
ribution offline. Second SAM encompasses diverse attention mechanisms (i.e. self
-attention and two-way cross-attention) resulting in substantial variations in t
he post-Softmax distributions. Therefore we introduce an Adaptive Granularity Qu
antization for Softmax through searching the optimal power-of-two base which is
hardware-friendly. Extensive experimental results across various vision tasks (i
nstance segmentation semantic segmentation and object detection) datasets and mo
del variants show the superiority of PTQ4SAM. For example when quantizing SAM-L
to 6-bit we achieve lossless accuracy for instance segmentation about 0.5% drop
with theoretical 3.9xacceleration. The code is available at https://github.com/c
hengtao-lv/PTQ4SAM.
****
Improving Bird's Eye View Semantic Segmentation by Task Decomposition
Tianhao Zhao, Yongcan Chen, Yu Wu, Tianyang Liu, Bo Du, Peilun Xiao, Shi Qiu, Ho
ngda Yang, Guozhen Li, Yi Yang, Yutian Lin; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15512-15521
Semantic segmentation in bird's eye view (BEV) plays a crucial role in autonomou
s driving. Previous methods usually follow an end-to-end pipeline directly predi
cting the BEV segmentation map from monocular RGB inputs. However the challenge
arises when the RGB inputs and BEV targets from distinct perspectives making the
 direct point-to-point predicting hard to optimize. In this paper we decompose t
he original BEV segmentation task into two stages namely BEV map reconstruction
and RGB-BEV feature alignment. In the first stage we train a BEV autoencoder to
reconstruct the BEV segmentation maps given corrupted noisy latent representatio
n which urges the decoder to learn fundamental knowledge of typical BEV patterns
. The second stage involves mapping RGB input images into the BEV latent space o
f the first stage directly optimizing the correlations between the two views at
the feature level. Our approach simplifies the complexity of combining perceptio
n and generation into distinct steps equipping the model to handle intricate and
 challenging scenes effectively. Besides we propose to transform the BEV segment
ation map from the Cartesian to the polar coordinate system to establish the col

umn-wise correspondence between RGB images and BEV maps. Moreover our method req uires neither multi-scale features nor camera intrinsic parameters for depth est imation and saves computational overhead. Extensive experiments on nuScenes and Argoverse show the effectiveness and efficiency of our method. Code is available at https://github.com/happytianhao/TaDe.
****
SpikingResformer: Bridging ResNet and Vision Transformer in Spiking Neural Netwo rks
Xinyu Shi, Zecheng Hao, Zhaofei Yu; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2024, pp. 5610-5619
The remarkable success of Vision Transformers in Artificial Neural Networks (ANN s) has led to a growing interest in incorporating the self-attention mechanism a nd transformer-based architecture into Spiking Neural Networks (SNNs). While exi sting methods propose spiking self-attention mechanisms that are compatible with SNNs they lack reasonable scaling methods and the overall architectures propose d by these methods suffer from a bottleneck in effectively extracting local feat ures. To address these challenges we propose a novel spiking self-attention mech anism named Dual Spike Self-Attention (DSSA) with a reasonable scaling method. B ased on DSSA we propose a novel spiking Vision Transformer architecture called S pikingResformer which combines the ResNet-based multi-stage architecture with ou r proposed DSSA to improve both performance and energy efficiency while reducing parameters. Experimental results show that SpikingResformer achieves higher acc uracy with fewer parameters and lower energy consumption than other spiking Visi on Transformer counterparts. Notably our SpikingResformer-L achieves 79.40% top-1 accuracy on ImageNet with 4 time-steps which is the state-of-the-art result in the SNN field.
****
Scene Adaptive Sparse Transformer for Event-based Object Detection
Yansong Peng, Hebei Li, Yueyi Zhang, Xiaoyan Sun, Feng Wu; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1 6794-16804
While recent Transformer-based approaches have shown impressive performances on event-based object detection tasks their high computational costs still diminish the low power consumption advantage of event cameras. Image-based works attempt to reduce these costs by introducing sparse Transformers. However they display inadequate sparsity and adaptability when applied to event-based object detectio n since these approaches cannot balance the fine granularity of token-level spar sification and the efficiency of window-based Transformers leading to reduced pe rformance and efficiency. Furthermore they lack scene-specific sparsity optimiza tion resulting in information loss and a lower recall rate. To overcome these li mitations we propose the Scene Adaptive Sparse Transformer (SAST). SAST enables window-token co-sparsification significantly enhancing fault tolerance and reduc ing computational overhead. Leveraging the innovative scoring and selection modu les along with the Masked Sparse Window Self-Attention SAST showcases remarkable scene-aware adaptability: It focuses only on important objects and dynamically optimizes sparsity level according to scene complexity maintaining a remarkable balance between performance and computational cost. The evaluation results show that SAST outperforms all other dense and sparse networks in both performance an d efficiency on two large-scale event-based object detection datasets (1Mpx and Gen1). Code: https://github.com/Peterande/SAST
****
Gaussian Shadow Casting for Neural Characters
Luis Bolanos, Shih-Yang Su, Helge Rhodin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20997-21006
Neural character models can now reconstruct detailed geometry and texture from v ideo but they lack explicit shadows and shading leading to artifacts when genera ting novel views and poses or during relighting. It is particularly difficult to include shadows as they are a global effect and the required casting of seconda ry rays is costly. We propose a new shadow model using a Gaussian density proxy that replaces sampling with a simple analytic formula. It supports dynamic motio

n and is tailored for shadow computation thereby avoiding the affine projection approximation and sorting required by the closely related Gaussian splatting. Combined with a deferred neural rendering model our Gaussian shadows enable Lambertian shading and shadow casting with minimal overhead. We demonstrate improved reconstructions with better separation of albedo shading and shadows in challenging outdoor scenes with direct sun light and hard shadows. Our method is able to optimize the light direction without any input from the user. As a result novel poses have fewer shadow artifacts and relighting in novel scenes is more realistic compared to the state-of-the-art methods providing new ways to pose neural characters in novel environments increasing their applicability. Code available at: https://github.com/LuisBolanos17/GaussianShadowCasting
****

CURSOR: Scalable Mixed-Order Hypergraph Matching with CUR Decomposition
Qixuan Zheng, Ming Zhang, Hong Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16036-16045
To achieve greater accuracy hypergraph matching algorithms require exponential increases in computational resources. Recent kd-tree-based approximate nearest neighbor (ANN) methods despite the sparsity of their compatibility tensor still require exhaustive calculations for large-scale graph matching. This work utilizes CUR tensor decomposition and introduces a novel cascaded second and third-order hypergraph matching framework (CURSOR) for efficient hypergraph matching. A CUR-based second-order graph matching algorithm is used to provide a rough match and then the core of CURSOR a fiber-CUR-based tensor generation method directly calculates entries of the compatibility tensor by leveraging the initial second-order match result. This significantly decreases the time complexity and tensor density. A probability relaxation labeling (PRL)-based matching algorithm especially suitable for sparse tensors is developed. Experiment results on large-scale synthetic datasets and widely-adopted benchmark sets demonstrate the superiority of CURSOR over existing methods. The tensor generation method in CURSOR can be integrated seamlessly into existing hypergraph matching methods to improve their performance and lower their computational costs.
****

Federated Online Adaptation for Deep Stereo
Matteo Poggi, Fabio Tosi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20165-20175
We introduce a novel approach for adapting deep stereo networks in a collaborative manner. By building over principles of federated learning we develop a distributed framework allowing for demanding the optimization process to a number of clients deployed in different environments. This makes it possible for a deep stereo network running on resourced-constrained devices to capitalize on the adaptation process carried out by other instances of the same architecture and thus improve its accuracy in challenging environments even when it cannot carry out adaptation on its own. Experimental results show how federated adaptation performs equivalently to on-device adaptation and even better when dealing with challenging environments.
****

Sequential Modeling Enables Scalable Learning for Large Vision Models
Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L. Yuille, Trevor Darrell, Jitendra Malik, Alexei A. Efros; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22861-22872
We introduce a novel sequential modeling approach which enables learning a Large Vision Model (LVM) without making use of any linguistic data. To do this we define a common format "visual sentences" in which we can represent raw images and videos as well as annotated data sources such as semantic segmentations and depth reconstructions without needing any meta-knowledge beyond the pixels. Once this wide variety of visual data (comprising 420 billion tokens) is represented as sequences the model can be trained to minimize a cross-entropy loss for next token prediction. By training across various scales of model architecture and data diversity we provide empirical evidence that our models scale effectively. Many different vision tasks can be solved by designing suitable visual prompts at tes

t time.
****

Self-Supervised Dual Contouring
Ramana Sundararaman, Roman Klokov, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4681-4691
Learning-based isosurface extraction methods have recently emerged as a robust and efficient alternative to axiomatic techniques. However the vast majority of such approaches rely on supervised training with axiomatically computed ground truths thus potentially inheriting biases and data artefacts of the corresponding axiomatic methods. Steering away from such dependencies we propose a self-supervised training scheme to the Neural Dual Contouring meshing framework resulting in our method: Self-Supervised Dual Contouring (SDC). Instead of optimizing predicted mesh vertices with supervised training we use two novel self-supervised loss functions that encourage the consistency between distances to the generated mesh up to the first order. Meshes reconstructed by SDC surpass existing data-driven methods in capturing intricate details while being more robust to possible irregularities in the input. Furthermore we use the same self-supervised training objective linking inferred mesh and input SDF to regularize the training process of Deep Implicit Networks (DINs). We demonstrate that the resulting DINs produce higher-quality implicit functions ultimately leading to more accurate and detail-preserving surfaces compared to prior baselines for different input modalities. Finally we demonstrate that our self-supervised losses improve meshing performance in the single-view reconstruction task by enabling joint training of predicted SDF and resulting output mesh.
****

Regularized Parameter Uncertainty for Improving Generalization in Reinforcement Learning
Pehuen Moure, Longbiao Cheng, Joachim Ott, Zuowen Wang, Shih-Chii Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23805-23814
In order for reinforcement learning (RL) agents to be deployed in real-world environments they must be able to generalize to unseen environments. However RL struggles with out-of-distribution generalization often due to over-fitting the particulars of the training environment. Although regularization techniques from supervised learning can be applied to avoid over-fitting the differences between supervised learning and RL limit their application. To address this we propose the Signal-to-Noise Ratio regulated Parameter Uncertainty Network (SNR PUN) for RL. We introduce SNR as a new measure of regularizing the parameter uncertainty of a network and provide a formal analysis explaining why SNR regularization works well for RL. We demonstrate the effectiveness of our proposed method to generalize in several simulated environments; and in a physical system showing the possibility of using SNR PUN for applying RL to real-world applications.
****

GigaTraj: Predicting Long-term Trajectories of Hundreds of Pedestrians in Gigapixel Complex Scenes
Haozhe Lin, Chunyu Wei, Li He, Yuchen Guo, Yunqi Zhao, Shanglong Li, Lu Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19331-19340
Pedestrian trajectory prediction is a well-established task with significant recent advancements. However existing datasets are unable to fulfill the demand for studying minute-level long-term trajectory prediction mainly due to the lack of high-resolution trajectory observation in the wide field of view (FoV). To bridge this gap we introduce a novel dataset named GigaTraj featuring videos capturing a wide FoV with ~ 4 x10^4 m^2 and high-resolution imagery at the gigapixel level. Furthermore GigaTraj includes comprehensive annotations such as bounding boxes identity associations world coordinates group/interaction relationships and scene semantics. Leveraging these multimodal annotations we evaluate and validate the state-of-the-art approaches for minute-level long-term trajectory prediction in large-scale scenes. Extensive experiments and analyses have revealed that

long-term prediction for pedestrian trajectories presents numerous challenges i
ndicating a vital new direction for trajectory research. The dataset is availabl
e at www.gigavision.ai.
****
GSVA: Generalized Segmentation via Multimodal Large Language Models
Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, Gao Huang; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 3858-3869
Generalized Referring Expression Segmentation (GRES) extends the scope of classi
c RES to refer to multiple objects in one expression or identify the empty targe
ts absent in the image. GRES poses challenges in modeling the complex spatial re
lationships of the instances in the image and identifying non-existing referents
. Multimodal Large Language Models (MLLMs) have recently shown tremendous progre
ss in these complicated vision-language tasks. Connecting Large Language Models
(LLMs) and vision models MLLMs are proficient in understanding contexts with vis
ual inputs. Among them LISA as a representative adopts a special [SEG] token to
prompt a segmentation mask decoder e.g. SAM to enable MLLMs in the RES task. How
ever existing solutions to GRES remain unsatisfactory since current segmentation
 MLLMs cannot correctly handle the cases where users might reference multiple su
bjects in a singular prompt or provide descriptions incongruent with any image t
arget. In this paper we propose Generalized Segmentation Vision Assistant (GSVA)
 to address this gap. Specifically GSVA reuses the [SEG] token to prompt the seg
mentation model towards supporting multiple mask references simultaneously and i
nnovatively learns to generate a [REJ] token to reject the targets explicitly. E
xperiments validate GSVA's efficacy in resolving the GRES issue marking a notabl
e enhancement and setting a new record on the GRES benchmark gRefCOCO dataset. G
SVA also proves effective across various classic referring segmentation and comp
rehension tasks.
****
AdaBM: On-the-Fly Adaptive Bit Mapping for Image Super-Resolution
Cheeun Hong, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2024, pp. 2641-2650
Although image super-resolution (SR) problem has experienced unprecedented resto
ration accuracy with deep neural networks it has yet limited versatile applicati
ons due to the substantial computational costs. Since different input images for
 SR face different restoration difficulties adapting computational costs based o
n the input image referred to as adaptive inference has emerged as a promising s
olution to compress SR networks. Specifically adapting the quantization bit-widt
hs has successfully reduced the inference and memory cost without sacrificing th
e accuracy. However despite the benefits of the resultant adaptive network exist
ing works rely on time-intensive quantization-aware training with full access to
 the original training pairs to learn the appropriate bit allocation policies wh
ich limits its ubiquitous usage. To this end we introduce the first on-the-fly a
daptive quantization framework that accelerates the processing time from hours t
o seconds. We formulate the bit allocation problem with only two bit mapping mod
ules: one to map the input image to the image-wise bit adaptation factor and one
 to obtain the layer-wise adaptation factors. These bit mappings are calibrated
and fine-tuned using only a small number of calibration images. We achieve compe
titive performance with the previous adaptive quantization methods while the pro
cessing time is accelerated by x2000. Codes are available at https://github.com/
Cheeun/AdaBM.
****
CoralSCOP: Segment any COral Image on this Planet
Ziqiang Zheng, Haixin Liang, Binh-Son Hua, Yue Him Wong, Put Ang Jr, Apple Pui Y
i Chui, Sai-Kit Yeung; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2024, pp. 28170-28180
Underwater visual understanding has recently gained increasing attention within
the computer vision community for studying and monitoring underwater ecosystems.
 Among these coral reefs play an important and intricate role often referred to
as the rainforests of the sea due to their rich biodiversity and crucial environ

mental impact. Existing coral analysis due to its technical complexity requires significant manual work from coral biologists therefore hindering scalable and comprehensive studies. In this paper we introduce CoralSCOP the first foundation model designed for the automatic dense segmentation of coral reefs. CoralSCOP is developed to accurately assign labels to different coral entities addressing the challenges in the semantic analysis of coral imagery. Its main objective is to identify and delineate the irregular boundaries between various coral individuals across different granularities such as coral/non-coral growth form and genus. This task is challenging due to the semantic agnostic nature or fixed limited semantic categories of previous generic segmentation methods which fail to adequately capture the complex characteristics of coral structures. By introducing a novel parallel semantic branch CoralSCOP can produce high-quality coral masks with semantics that enable a wide range of downstream coral reef analysis tasks. We demonstrate that CoralSCOP exhibits a strong zero-shot ability to segment unseen coral images. To effectively train our foundation model we propose CoralMask a new dataset with 41297 densely labeled coral images and 330144 coral masks. We have conducted comprehensive and extensive experiments to demonstrate the advantages of CoralSCOP over existing generalist segmentation algorithms and coral reef analytical approaches.
****
SVGDreamer: Text Guided SVG Generation with Diffusion Model
Ximing Xing, Haitao Zhou, Chuang Wang, Jing Zhang, Dong Xu, Qian Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4546-4555
Recently text-guided scalable vector graphics (SVGs) synthesis has shown promise in domains such as iconography and sketch. However existing text-to-SVG generation methods lack editability and struggle with visual quality and result diversity. To address these limitations we propose a novel text-guided vector graphics synthesis method called SVGDreamer. SVGDreamer incorporates a semantic-driven image vectorization (SIVE) process that enables the decomposition of synthesis into foreground objects and background thereby enhancing editability. Specifically the SIVE process introduces attention-based primitive control and an attention-mask loss function for effective control and manipulation of individual elements. Additionally we propose a Vectorized Particle-based Score Distillation (VPSD) approach to address issues of shape over-smoothing color over-saturation limited diversity and slow convergence of the existing text-to-SVG generation methods by modeling SVGs as distributions of control points and colors. Furthermore VPSD leverages a reward model to re-weight vector particles which improves aesthetic appeal and accelerates convergence. Extensive experiments are conducted to validate the effectiveness of SVGDreamer demonstrating its superiority over baseline methods in terms of editability visual quality and diversity. Project page: \href https://ximinng.github.io/SVGDreamer-project/ https://ximinng.github.io/SVGDreamer-project/
****
BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition
Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2049-2058
Graph Convolutional Networks (GCNs) have long set the state-of-the-art in skeleton-based action recognition leveraging their ability to unravel the complex dynamics of human joint topology through the graph's adjacency matrix. However an inherent flaw has come to light in these cutting-edge models: they tend to optimize the adjacency matrix jointly with the model weights. This process while seemingly efficient causes a gradual decay of bone connectivity data resulting in a model indifferent to the very topology it sought to represent. To remedy this we propose a two-fold strategy: (1) We introduce an innovative approach that encodes bone connectivity by harnessing the power of graph distances to describe the physical topology; we further incorporate action-specific topological representation via persistent homology analysis to depict systemic dynamics. This preserves the vital topological nuances often lost in conventional GCNs. (2) Our investiga

tion also reveals the redundancy in existing GCNs for multi-relational modeling which we address by proposing an efficient refinement to Graph Convolutions (GC) - the BlockGC. This significantly reduces parameters while improving performance beyond original GCNs. Our full model BlockGCN establishes new benchmarks in skeleton-based action recognition across all model categories. Its high accuracy and lightweight design most notably on the large-scale NTU RGB+D 120 dataset stand as strong validation of the efficacy of BlockGCN.
****
Improved Baselines with Visual Instruction Tuning
Haotian Liu, Chunyuan Li, Yuheng Li, Yong Jae Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26296-26306
Large multimodal models (LMM) have recently shown encouraging progress with visual instruction tuning. In this paper we present the first systematic study to investigate the design choices of LMMs in a controlled setting under the LLaVA framework. We show that the fully-connected vision-language connector in LLaVA is surprisingly powerful and data-efficient. With simple modifications to LLaVA namely using CLIP-ViT-L-336px with an MLP projection and adding academic-task-oriented VQA data with response formatting prompts we establish stronger baselines that achieve state-of-the-art across 11 benchmarks. Our final 13B checkpoint uses merely 1.2M publicly available data and finishes full training in 1 day on a single 8-A100 node. Furthermore we present some early exploration of open problems in LMMs including scaling to higher resolution inputs compositional capabilities and model hallucination etc. We hope this makes state-of-the-art LMM research more accessible. Code and model will be publicly available.
****
Structure-Guided Adversarial Training of Diffusion Models
Ling Yang, Haotian Qian, Zhilong Zhang, Jingwei Liu, Bin Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7256-7266
Diffusion models have demonstrated exceptional efficacy in various generative applications. While existing models focus on minimizing a weighted sum of denoising score matching losses for data distribution modeling their training primarily emphasizes instance-level optimization overlooking valuable structural information within each mini-batch indicative of pair-wise relationships among samples. To address this limitation we introduce Structure-guided Adversarial training of Diffusion Models (SADM). In this pioneering approach we compel the model to learn manifold structures between samples in each training batch. To ensure the model captures authentic manifold structures in the data distribution we advocate adversarial training of the diffusion generator against a novel structure discriminator in a minimax game distinguishing real manifold structures from the generated ones. SADM substantially outperforms existing methods in image generation and cross-domain fine-tuning tasks across 12 datasets establishing a new state-of-the-art FID of 1.58 and 2.11 on ImageNet for class-conditional image generation at resolutions of 256x256 and 512x512 respectively.
****
NIFTY: Neural Object Interaction Fields for Guided Human Motion Synthesis
Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, Leonidas Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 947-957
We address the problem of generating realistic 3D motions of humans interacting with objects in a scene. Our key idea is to create a neural interaction field attached to a specific object which outputs the distance to the valid interaction manifold given a human pose as input. This interaction field guides the sampling of an object-conditioned human motion diffusion model so as to encourage plausible contacts and affordance semantics. To support interactions with scarcely available data we propose an automated synthetic data pipeline. For this we seed a pre-trained motion model which has priors for the basics of human movement with interaction-specific anchor poses extracted from limited motion capture data. Using our guided diffusion model trained on generated synthetic data we synthesize

realistic motions for sitting and lifting with several objects outperforming alternative approaches in terms of motion quality and successful action completion. We call our framework NIFTY: Neural Interaction Fields for Trajectory sYnthesis.
****
C2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation
Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, Song Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16006-16015
Existing Knowledge Distillation (KD) methods typically focus on transferring knowledge from a large-capacity teacher to a low-capacity student model achieving substantial success in unimodal knowledge transfer. However existing methods can hardly be extended to Cross-Modal Knowledge Distillation (CMKD) where the knowledge is transferred from a teacher modality to a different student modality with inference only on the distilled student modality. We empirically reveal that the modality gap i.e. modality imbalance and soft label misalignment incurs the ineffectiveness of traditional KD in CMKD. As a solution we propose a novel \underline C ustomized \underline C rossmodal \underline K nowledge \underline D istillation (C^2KD). Specifically to alleviate the modality gap the pre-trained teacher performs bidirectional distillation with the student to provide customized knowledge. The On-the-Fly Selection Distillation(OFSD) strategy is applied to selectively filter out the samples with misaligned soft labels where we distill cross-modal knowledge from non-target classes to avoid the modality imbalance issue. To further provide receptive cross-modal knowledge proxy student and teacher inheriting unimodal and cross-modal knowledge is formulated to progressively transfer cross-modal knowledge through bidirectional distillation. Experimental results on audio-visual image-text and RGB-depth datasets demonstrate that our method can effectively transfer knowledge across modalities achieving superior performance against traditional KD by a large margin.
****
Traceable Federated Continual Learning
Qiang Wang, Bingyan Liu, Yawen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12872-12881
Federated continual learning (FCL) is a typical mechanism to achieve collaborative model training among clients that own dynamic data. While traditional FCL methods have been proved effective they do not consider the task repeatability and fail to achieve good performance under this practical scenario. In this paper we propose a new paradigm namely Traceable Federated Continual Learning (TFCL) aiming to cope with repetitive tasks by tracing and augmenting them. Following the new paradigm we develop TagFed a framework that enables accurate and effective Tracing augmentation and Federation for TFCL. The key idea is to decompose the whole model into a series of marked sub-models for optimizing each client task before conducting group-wise knowledge aggregation such that the repetitive tasks can be located precisely and federated selectively for improved performance. Extensive experiments on our constructed benchmark demonstrate the effectiveness and efficiency of the proposed framework. We will release our code at: https://github.com/P0werWeirdo/TagFCL.
****
Can Language Beat Numerical Regression? Language-Based Multimodal Trajectory Prediction
Inhwan Bae, Junoh Lee, Hae-Gon Jeon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 753-766
Language models have demonstrated impressive ability in context understanding and generative performance. Inspired by the recent success of language foundation models in this paper we propose LMTraj (Language-based Multimodal Trajectory predictor) which recasts the trajectory prediction task into a sort of question-answering problem. Departing from traditional numerical regression models which treat the trajectory coordinate sequence as continuous signals we consider them as discrete signals like text prompts. Specially we first transform an input space for the trajectory coordinate into the natural language space. Here the entire t

ime-series trajectories of pedestrians are converted into a text prompt and scen e images are described as text information through image captioning. The transfo rmed numerical and image data are then wrapped into the question-answering templ ate for use in a language model. Next to guide the language model in understandi ng and reasoning high-level knowledge such as scene context and social relations hips between pedestrians we introduce an auxiliary multi-task question and answe ring. We then train a numerical tokenizer with the prompt data. We encourage the tokenizer to separate the integer and decimal parts well and leverage it to cap ture correlations between the consecutive numbers in the language model. Lastly we train the language model using the numerical tokenizer and all of the questio n-answer prompts. Here we propose a beam-search-based most-likely prediction and a temperature-based multimodal prediction to implement both deterministic and s tochastic inferences. Applying our LMTraj we show that the language-based model can be a powerful pedestrian trajectory predictor and outperforms existing numer ical-based predictor methods. Extensive experiments show that our LMTraj can suc cessfully understand social relationships and accurately extrapolate the multimo dal futures on the public pedestrian trajectory prediction benchmark. Code is pu blicly available at https://github.com/inhwanbae/LMTrajectory.
****
Building Optimal Neural Architectures using Interpretable Knowledge
Keith G. Mills, Fred X. Han, Mohammad Salameh, Shengyao Lu, Chunhua Zhou, Jiao H e, Fengyu Sun, Di Niu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5726-5735
Neural Architecture Search is a costly practice. The fact that a search space ca n span a vast number of design choices with each architecture evaluation taking nontrivial overhead makes it hard for an algorithm to sufficiently explore candi date networks. In this paper we propose AutoBuild a scheme which learns to align the latent embeddings of operations and architecture modules with the ground-tr uth performance of the architectures they appear in. By doing so AutoBuild is ca pable of assigning interpretable importance scores to architecture modules such as individual operation features and larger macro operation sequences such that high-performance neural networks can be constructed without any need for search. Through experiments performed on state-of-the-art image classification segmenta tion and Stable Diffusion models we show that by mining a relatively small set o f evaluated architectures AutoBuild can learn to build high-quality architecture s directly or help to reduce search space to focus on relevant areas finding bet ter architectures that outperform both the original labeled ones and ones found by search baselines. Code available at https://github.com/Ascend-Research/AutoBu ild
****
V?: Guided Visual Search as a Core Mechanism in Multimodal LLMs
Penghao Wu, Saining Xie; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2024, pp. 13084-13094
When we look around and perform complex tasks how we see and selectively process what we see is crucial. However the lack of this visual search mechanism in cur rent multimodal LLMs (MLLMs) hinders their ability to focus on important visual details especially when handling high-resolution and visually crowded images. To address this we introduce V* an LLM-guided visual search mechanism that employs the world knowledge in LLMs for efficient visual querying. When combined with a n MLLM this mechanism enhances collaborative reasoning contextual understanding and precise visual grounding. This integration results in a new MLLM meta-archit ecture named Show sEArch and TelL (SEAL). We further create V*Bench a benchmark specifically designed to evaluate MLLMs in their ability to process high-resolut ion images and focus on visual details. Our study highlights the necessity of in corporating visual search capabilities into multimodal systems. The code is avai lable at https://github.com/penghao-wu/vstar
****
Unexplored Faces of Robustness and Out-of-Distribution: Covariate Shifts in Envi ronment and Sensor Domains
Eunsu Baek, Keondo Park, Jiyoon Kim, Hyung-Sin Kim; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22294-22303
Computer vision applications predict on digital images acquired by a camera from physical scenes through light. However conventional robustness benchmarks rely on perturbations in digitized images diverging from distribution shifts occurring in the image acquisition process. To bridge this gap we introduce a new distribution shift dataset ImageNet-ES comprising variations in environmental and camera sensor factors by directly capturing 202k images with a real camera in a controllable testbed. With the new dataset we evaluate out-of-distribution (OOD) detection and model robustness. We find that existing OOD detection methods do not cope with the covariate shifts in ImageNet-ES implying that the definition and detection of OOD should be revisited to embrace real-world distribution shifts. We also observe that the model becomes more robust in both ImageNet-C and -ES by learning environment and sensor variations in addition to existing digital augmentations. Lastly our results suggest that effective shift mitigation via camera sensor control can significantly improve performance without increasing model size. With these findings our benchmark may aid future research on robustness OOD and camera sensor control for computer vision. Our code and dataset are available at https://github.com/Edw2n/ImageNet-ES.
****
Uncertainty Visualization via Low-Dimensional Posterior Projections
Omer Yair, Elias Nehme, Tomer Michaeli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11041-11051
In ill-posed inverse problems it is commonly desirable to obtain insight into the full spectrum of plausible solutions rather than extracting only a single reconstruction. Information about the plausible solutions and their likelihoods is encoded in the posterior distribution. However for high-dimensional data this distribution is challenging to visualize. In this work we introduce a new approach for estimating and visualizing posteriors by employing energy-based models (EBMs) over low-dimensional subspaces. Specifically we train a conditional EBM that receives an input measurement and a set of directions that span some low-dimensional subspace of solutions and outputs the probability density function of the posterior within that space. We demonstrate the effectiveness of our method across a diverse range of datasets and image restoration problems showcasing its strength in uncertainty quantification and visualization. As we show our method outperforms a baseline that projects samples from a diffusion-based posterior sampler while being orders of magnitude faster. Furthermore it is more accurate than a baseline that assumes a Gaussian posterior.
****
VSCode: General Visual Salient and Camouflaged Object Detection with 2D Prompt Learning
Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, Junwei Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17169-17180
Salient object detection (SOD) and camouflaged object detection (COD) are related yet distinct binary mapping tasks. These tasks involve multiple modalities sharing commonalities and unique cues. Existing research often employs intricate task-specific specialist models potentially leading to redundancy and suboptimal results. We introduce VSCode a generalist model with novel 2D prompt learning to jointly address four SOD tasks and three COD tasks. We utilize VST as the foundation model and introduce 2D prompts within the encoder-decoder architecture to learn domain and task-specific knowledge on two separate dimensions. A prompt discrimination loss helps disentangle peculiarities to benefit model optimization. VSCode outperforms state-of-the-art methods across six tasks on 26 datasets and exhibits zero-shot generalization to unseen tasks by combining 2D prompts such as RGB-D COD. Source code has been available at https://github.com/Sssssuperior/VSCode.
****
GaussianEditor: Swift and Controllable 3D Editing with Gaussian Splatting
Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhonga

ng Cai, Lei Yang, Huaping Liu, Guosheng Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21476-21485
3D editing plays a crucial role in many areas such as gaming and virtual reality. Traditional 3D editing methods which rely on representations like meshes and point clouds often fall short in realistically depicting complex scenes. On the other hand methods based on implicit 3D representations like Neural Radiance Field (NeRF) render complex scenes effectively but suffer from slow processing speeds and limited control over specific scene areas. In response to these challenges our paper presents GaussianEditor the first 3D editing algorithm based on Gaussian Splatting (GS) a novel 3D representation. GaussianEditor enhances precision and control in editing through our proposed Gaussian semantic tracing which traces the editing target throughout the training process. Additionally we propose Hierarchical Gaussian splatting (HGS) to achieve stabilized and fine results under stochastic generative guidance from 2D diffusion models. We also develop editing strategies for efficient object removal and integration a challenging task for existing methods. Our comprehensive experiments demonstrate GaussianEditor's superior control effective and efficient performance marking a significant advancement in 3D editing.
****
Holo-Relighting: Controllable Volumetric Portrait Relighting from a Single Image
Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4263-4273
At the core of portrait photography is the search for ideal lighting and viewpoint. The process often requires advanced knowledge in photography and an elaborate studio setup. In this work we propose Holo-Relighting a volumetric relighting method that is capable of synthesizing novel viewpoints and novel lighting from a single image. Holo-Relighting leverages the pretrained 3D GAN (EG3D) to reconstruct geometry and appearance from an input portrait as a set of 3D-aware features. We design a relighting module conditioned on a given lighting to process these features and predict a relit 3D representation in the form of a tri-plane which can render to an arbitrary viewpoint through volume rendering. Besides viewpoint and lighting control Holo-Relighting also takes the head pose as a condition to enable head-pose-dependent lighting effects. With these novel designs Holo-Relighting can generate complex non-Lambertian lighting effects (e.g. specular highlights and cast shadows) without using any explicit physical lighting priors. We train Holo-Relighting with data captured with a light stage and propose two data-rendering techniques to improve the data quality for training the volumetric relighting system. Through quantitative and qualitative experiments we demonstrate Holo-Relighting can achieve state-of-the-arts relighting quality with better photorealism 3D consistency and controllability.
****
Noisy One-point Homographies are Surprisingly Good
Yaqing Ding, Jonathan Astermark, Magnus Oskarsson, Viktor Larsson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5125-5134
Two-view homography estimation is a classic and fundamental problem in computer vision. While conceptually simple the problem quickly becomes challenging when multiple planes are visible in the image pair. Even with correct matches each individual plane (homography) might have a very low number of inliers when comparing to the set of all correspondences. In practice this requires a large number of RANSAC iterations to generate a good model hypothesis. The current state-of-the-art methods therefore seek to reduce the sample size from four point correspondences originally by including additional information such as keypoint orientation/angles or local affine information. In this work we continue in this direction and propose a novel one-point solver that leverages different approximate constraints derived from the same auxiliary information. In experiments we obtain state-of-the-art results with execution time speed-ups on large benchmark datasets and show that it is more beneficial for the solver to be sample efficient compared to generating more accurate homographies.

****
PointInfinity: Resolution-Invariant Point Diffusion Models
Zixuan Huang, Justin Johnson, Shoubhik Debnath, James M. Rehg, Chao-Yuan Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10050-10060
We present PointInfinity an efficient family of point cloud diffusion models. Our core idea is to use a transformer-based architecture with a fixed-size resolution-invariant latent representation. This enables efficient training with low-resolution point clouds while allowing high-resolution point clouds to be generated during inference. More importantly we show that scaling the test-time resolution beyond the training resolution improves the fidelity of generated point clouds and surfaces. We analyze this phenomenon and draw a link to classifier-free guidance commonly used in diffusion models demonstrating that both allow trading off fidelity and variability during inference. Experiments on CO3D show that PointInfinity can efficiently generate high-resolution point clouds (up to 131k points 31 times more than Point-E) with state-of-the-art quality.
****
Panacea: Panoramic and Controllable Video Generation for Autonomous Driving
Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6902-6912
The field of autonomous driving increasingly demands high-quality annotated training data. In this paper we propose Panacea an innovative approach to generate panoramic and controllable videos in driving scenarios capable of yielding an unlimited numbers of diverse annotated samples pivotal for autonomous driving advancements. Panacea addresses two critical challenges: 'Consistency' and 'Controllability.' Consistency ensures temporal and cross-view coherence while Controllability ensures the alignment of generated content with corresponding annotations. Our approach integrates a novel 4D attention and a two-stage generation pipeline to maintain coherence supplemented by the ControlNet framework for meticulous control by the Bird's-Eye-View (BEV) layouts. Extensive qualitative and quantitative evaluations of Panacea on the nuScenes dataset prove its effectiveness in generating high-quality multi-view driving-scene videos. This work notably propels the field of autonomous driving by effectively augmenting the training dataset used for advanced BEV perception techniques.
****
Open-Vocabulary Semantic Segmentation with Image Embedding Balancing
Xiangheng Shan, Dongyue Wu, Guilin Zhu, Yuanjie Shao, Nong Sang, Changxin Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28412-28421
Open-vocabulary semantic segmentation is a challenging task which requires the model to output semantic masks of an image beyond a close-set vocabulary. Although many efforts have been made to utilize powerful CLIP models to accomplish this task they are still easily overfitting to training classes due to the natural gaps in semantic information between training and new classes. To overcome this challenge we propose a novel framework for open-vocabulary semantic segmentation called EBSeg incorporating an Adaptively Balanced Decoder (AdaB Decoder) and a Semantic Structure Consistency loss (SSC Loss). The AdaB Decoder is designed to generate different image embeddings for both training and new classes. Subsequently these two types of embeddings are adaptively balanced to fully exploit their ability to recognize training classes and generalization ability for new classes. To learn a consistent semantic structure from CLIP the SSC Loss aligns the inter-classes affinity in the image feature space with that in the text feature space of CLIP thereby improving the generalization ability of our model. Furthermore we employ a frozen SAM image encoder to complement the spatial information that CLIP features lack due to the low training image resolution and image-level supervision inherent in CLIP. Extensive experiments conducted across various benchmarks demonstrate that the proposed EBSeg outperforms the state-of-the-art methods. Our code and trained models will be here: https://github.com/slonetime/EBSeg.

****
Structured Model Probing: Empowering Efficient Transfer Learning by Structured R
egularization
Zhi-Fan Wu, Chaojie Mao, Wue Wang, Jianwen Jiang, Yiliang Lv, Rong Jin; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2024, pp. 16849-16858
Despite encouraging results from recent developments in transfer learning for ad
apting pre-trained model to downstream tasks the performance of model probing is
 still lagging behind the state-of-the-art parameter efficient tuning methods. O
ur investigation reveals that existing model probing methods perform well for th
e easy case when the source domain (where models are pre-trained) and the adapte
d domain are similar but fail for the difficult case when the two domains are si
gnificantly different. Simply incorporating features extracted from multiple lay
ers and increasing complexity of the probing model can mitigate the gap in the d
ifficult case but degrades the performance in the easy case. To address this cha
llenge we propose structured model probing (SMP) that is able to deliver good pe
rformance for both cases through structured regularization. The regularization p
erforms feature selection leveraging model structure as a prior and controls the
 complexity of the probing model through the weights of selected structures. Thi
s enables us to construct a simple adaptation model with a small number of selec
ted features and a linear prediction model for the easy case; and to automatical
ly increase the complexity of adaptation model with a large number of selected f
eatures and a non-linear model for the difficult case. Our extensive empirical s
tudies show that SMP significantly outperforms the state-of-the-art methods for
parameter efficient tuning and at the same time still maintains the advantage of
 computational efficiency for probing-based methods.
****
Multi-Modal Proxy Learning Towards Personalized Visual Multiple Clustering
Jiawei Yao, Qi Qian, Juhua Hu; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2024, pp. 14066-14075
Multiple clustering has gained significant attention in recent years due to its
potential to reveal multiple hidden structures of data from different perspectiv
es. The advent of deep multiple clustering techniques has notably advanced the p
erformance by uncovering complex patterns and relationships within large dataset
s. However a major challenge arises as users often do not need all the clusterin
gs that algorithms generate and figuring out the one needed requires a substanti
al understanding of each clustering result. Traditionally aligning a user's brie
f keyword of interest with the corresponding vision components was challenging b
ut the emergence of multi-modal and large language models (LLMs) has begun to br
idge this gap. In response given unlabeled target visual data we propose Multi-M
ap a novel method employing a multi-modal proxy learning process. It leverages C
LIP encoders to extract coherent text and image embeddings with GPT-4 integratin
g users' interests to formulate effective textual contexts. Moreover reference w
ord constraint and concept-level constraint are designed to learn the optimal te
xt proxy according to the user's interest. Multi-Map not only adeptly captures a
 user's interest via a keyword but also facilitates identifying relevant cluster
ings. Our extensive experiments show that Multi-Map consistently outperforms sta
te-of-the-art methods in all benchmark multi-clustering vision tasks. Our code i
s available at https://github.com/Alexander-Yao/Multi-MaP.
****
DreamMatcher: Appearance Matching Self-Attention for Semantically-Consistent Tex
t-to-Image Personalization
Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, Seunggyu Chang; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2024, pp. 8100-8110
The objective of text-to-image (T2I) personalization is to customize a diffusion
 model to a user-provided reference concept generating diverse images of the con
cept aligned with the target prompts. Conventional methods representing the refe
rence concepts using unique text embeddings often fail to accurately mimic the a
ppearance of the reference. To address this one solution may be explicitly condi

tioning the reference images into the target denoising process known as key-valu e replacement. However prior works are constrained to local editing since they d isrupt the structure path of the pre-trained T2I model. To overcome this we prop ose a novel plug-in method called DreamMatcher which reformulates T2I personaliz ation as semantic matching. Specifically DreamMatcher replaces the target values with reference values aligned by semantic matching while leaving the structure path unchanged to preserve the versatile capability of pre-trained T2I models fo r generating diverse structures. We also introduce a semantic-consistent masking strategy to isolate the personalized concept from irrelevant regions introduced by the target prompts. Compatible with existing T2I models DreamMatcher shows s ignificant improvements in complex scenarios. Intensive analyses demonstrate the effectiveness of our approach.
****
Stronger Fewer & Superior: Harnessing Vision Foundation Models for Domain Genera lized Semantic Segmentation
Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang , Huaian Chen, Jinjin Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28619-28630
In this paper we first assess and harness various Vision Foundation Models (VFMs) in the context of Domain Generalized Semantic Segmentation (DGSS). Driven by t he motivation that Leveraging Stronger pre-trained models and Fewer trainable pa rameters for Superior generalizability we introduce a robust fine-tuning approac h namely "Rein" to parameter-efficiently harness VFMs for DGSS. Built upon a set of trainable tokens each linked to distinct instances Rein precisely refines an d forwards the feature maps from each layer to the next layer within the backbon e. This process produces diverse refinements for different categories within a s ingle image. With fewer trainable parameters Rein efficiently fine-tunes VFMs fo r DGSS tasks surprisingly surpassing full parameter fine-tuning. Extensive exper iments across various settings demonstrate that Rein significantly outperforms s tate-of-the-art methods. Remarkably with just an extra 1% of trainable parameter s within the frozen backbone Rein achieves a mIoU of 68.1% on the Cityscapes wit hout accessing any real urban-scene datasets. Code is available at https://githu b.com/w1oves/Rein.git.
****
PolarMatte: Fully Computational Ground-Truth-Quality Alpha Matte Extraction for Images and Video using Polarized Screen Matting
Kenji Enomoto, TJ Rhodes, Brian Price, Gavin Miller; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3901-39 09
The creation of high-quality alpha mattes as ground-truth data for video matting is typically a laborious task. The trade-off between accuracy manual correction s and capture constraints often produces erroneous results or is cost prohibitiv e. We propose PolarMatte a fully computational alpha matte extraction method for images and video without compromise between quality and practicality. A single polarization camera is used to capture dynamic scenes backlit by an off-the-shel f LCD monitor. PolarMatte exploits the polarization channel to compute the per-p ixel opacity of the target scene including the transparency of fine-details tran slucent objects and optical/motion blur. We leverage polarization clues to robus tly detect indistinguishable pixels and extract the alpha matte value at polariz ed foreground reflections with a polarimetric matting Laplacian. Quantitative an d qualitative evaluation demonstrate our ability to computationally extract grou nd-truth-quality alpha mattes without human labour.
****
ChAda-ViT : Channel Adaptive Attention for Joint Representation Learning of Hete rogeneous Microscopy Images
Nicolas Bourriez, Ihab Bendidi, Ethan Cohen, Gabriel Watkinson, Maxime Sanchez, Guillaume Bollot, Auguste Genovesio; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2024, pp. 11556-11565
Unlike color photography images which are consistently encoded into RGB channels biological images encompass various modalities where the type of microscopy and

the meaning of each channel varies with each experiment. Importantly the number of channels can range from one to a dozen and their correlation is often comparatively much lower than RGB as each of them brings specific information content. This aspect is largely overlooked by methods designed out of the bioimage field and current solutions mostly focus on intra-channel spatial attention often ignoring the relationship between channels yet crucial in most biological applications. Importantly the variable channel type and count prevent the projection of several experiments to a unified representation for large scale pre-training. In this study we propose ChAda-ViT a novel Channel Adaptive Vision Transformer architecture employing an Inter-Channel Attention mechanism on images with an arbitrary number order and type of channels. We also introduce IDRCell100k a bioimage dataset with a rich set of 79 experiments covering 7 microscope modalities with a multitude of channel types and channel counts varying from 1 to 10 per experiment. Our proposed architecture trained in a self-supervised manner outperforms existing approaches in several biologically relevant downstream tasks. Additionally it can be used to bridge the gap for the first time between assays with different microscopes channel numbers or types by embedding various image and experimental modalities into a unified biological image representation. The latter should facilitate interdisciplinary studies and pave the way for better adoption of deep learning in biological image-based analyses.
****
CARZero: Cross-Attention Alignment for Radiology Zero-Shot Classification
Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, S. Kevin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11137-11146
The advancement of Zero-Shot Learning in the medical domain has been driven forward by using pre-trained models on large-scale image-text pairs focusing on image-text alignment. However existing methods primarily rely on cosine similarity for alignment which may not fully capture the complex relationship between medical images and reports. To address this gap we introduce a novel approach called Cross-Attention Alignment for Radiology Zero-Shot Classification (CARZero). Our approach innovatively leverages cross-attention mechanisms to process image and report features creating a Similarity Representation that more accurately reflects the intricate relationships in medical semantics. This representation is then linearly projected to form an image-text similarity matrix for cross-modality alignment. Additionally recognizing the pivotal role of prompt selection in zero-shot learning CARZero incorporates a Large Language Model-based prompt alignment strategy. This strategy standardizes diverse diagnostic expressions into a unified format for both training and inference phases overcoming the challenges of manual prompt design. Our approach is simple yet effective demonstrating state-of-the-art performance in zero-shot classification on five official chest radiograph diagnostic test sets including remarkable results on datasets with long-tail distributions of rare diseases. This achievement is attributed to our new image-text alignment strategy which effectively addresses the complex relationship between medical images and reports. Code and models are available at https://github.com/laihaoran/CARZero.
****
HOIDiffusion: Generating Realistic 3D Hand-Object Interaction Data
Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8521-8531
3D hand-object interaction data is scarce due to the hardware constraints in scaling up the data collection process. In this paper we propose HOIDiffusion for generating realistic and diverse 3D hand-object interaction data. Our model is a conditional diffusion model that takes both the 3D hand-object geometric structure and text description as inputs for image synthesis. This offers a more controllable and realistic synthesis as we can specify the structure and style inputs in a disentangled manner. HOIDiffusion is trained by leveraging a diffusion model pre-trained on large-scale natural images and a few 3D human demonstrations. Beyond controllable image synthesis we adopt the generated 3D data for learning 6

D object pose estimation and show its effectiveness in improving perception systems. Project page: https://mq-zhang1.github.io/HOIDiffusion.
****
VecFusion: Vector Font Generation with Diffusion
Vikas Thamizharasan, Difan Liu, Shantanu Agarwal, Matthew Fisher, Michael Gharbi, Oliver Wang, Alec Jacobson, Evangelos Kalogerakis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7943-7952
We present VecFusion a new neural architecture that can generate vector fonts with varying topological structures and precise control point positions. Our approach is a cascaded diffusion model which consists of a raster diffusion model followed by a vector diffusion model. The raster model generates low-resolution rasterized fonts with auxiliary control point information capturing the global style and shape of the font while the vector model synthesizes vector fonts conditioned on the low-resolution raster fonts from the first stage. To synthesize long and complex curves our vector diffusion model uses a transformer architecture and a novel vector representation that enables the modeling of diverse vector geometry and the precise prediction of control points. Our experiments show that in contrast to previous generative models for vector graphics our new cascaded vector diffusion model generates higher quality vector fonts with complex structures and diverse styles.
****
Multi-Modal Hallucination Control by Visual Information Grounding
Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14303-14312
Generative Vision-Language Models (VLMs) are prone to generate plausible-sounding textual answers which however are not always grounded in the input image. We investigate this phenomenon usually referred to as "hallucination" and show that it stems from an excessive reliance on the language prior. In particular we show that as more tokens are generated the reliance on the visual prompt decreases and this behavior strongly correlates with the emergence of hallucinations. To reduce hallucinations we introduce Multi-Modal Mutual-Information Decoding (M3ID) a new sampling method for prompt amplification. M3ID amplifies the influence of the reference image over the language prior hence favoring the generation of tokens with higher mutual information with the visual prompt. M3ID can be applied to any pre-trained autoregressive VLM at inference time without necessitating further training and with minimal computational overhead. If training is an option we show that M3ID can be paired with Direct Preference Optimization (DPO) to improve the model's reliance on the prompt image without requiring any labels. Our empirical findings show that our algorithms maintain the fluency and linguistic capabilities of pre-trained VLMs while reducing hallucinations by mitigating visually ungrounded answers. Specifically for the LLaVA 13B model M3ID and M3ID+DPO reduce the percentage of hallucinated objects in captioning tasks by 25% and 28% respectively and improve the accuracy on VQA benchmarks such as POPE by 21% and 24%.
****
Towards Text-guided 3D Scene Composition
Qihang Zhang, Chaoyang Wang, Aliaksandr Siarohin, Peiye Zhuang, Yinghao Xu, Ceyuan Yang, Dahua Lin, Bolei Zhou, Sergey Tulyakov, Hsin-Ying Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6829-6838
We are witnessing significant breakthroughs in the technology for generating 3D objects from text. Existing approaches either leverage large text-to-image models to optimize a 3D representation or train 3D generators on object-centric datasets. Generating entire scenes however remains very challenging as a scene contains multiple 3D objects diverse and scattered. In this work we introduce SceneWiz3D - a novel approach to synthesize high-fidelity 3D scenes from text. We marry the locality of objects with globality of scenes by introducing a hybrid 3D repr

esentation - explicit for objects and implicit for scenes. Remarkably an object being represented explicitly can be either generated from text using conventiona l text-to-3D approaches or provided by users. To configure the layout of the sce ne and automatically place objects we apply the Particle Swarm Optimization tech nique during the optimization process. Furthermore it is difficult for certain p arts of the scene (e.g. corners occlusion) to receive multi-view supervision lea ding to inferior geometry. We incorporate an RGBD panorama diffusion model to mi tigate it resulting in high-quality geometry. Extensive evaluation supports that our approach achieves superior quality over previous approaches enabling the ge neration of detailed and view-consistent 3D scenes. Our project website is at ht tps://zqh0253.github.io/SceneWiz3D.\\
****

EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Mask ed Audio Gesture Modeling

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, X uefei Zhe, Naoya Iwamoto, Bo Zheng, Michael J. Black; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1144-1 154

We propose EMAGE a framework to generate full-body human gestures from audio and masked gestures encompassing facial local body hands and global movements. To a chieve this we first introduce BEAT2 (BEAT-SMPLX-FLAME) a new mesh-level holisti c co-speech dataset. BEAT2 combines a MoShed SMPL-X body with FLAME head paramet ers and further refines the modeling of head neck and finger movements offering a community-standardized high-quality 3D motion captured dataset. EMAGE leverage s masked body gesture priors during training to boost inference performance. It involves a Masked Audio Gesture Transformer facilitating joint training on audio -to-gesture generation and masked gesture reconstruction to effectively encode a udio and body gesture hints. Encoded body hints from masked gestures are then se parately employed to generate facial and body movements. Moreover EMAGE adaptive ly merges speech features from the audio's rhythm and content and utilizes four compositional VQ-VAEs to enhance the results' fidelity and diversity. Experiment s demonstrate that EMAGE generates holistic gestures with state-of-the-art perfo rmance and is flexible in accepting predefined spatial-temporal gesture inputs g enerating complete audio-synchronized results. Our code and dataset are availabl e. https://pantomatrix.github.io/EMAGE/
****

Adversarial Text to Continuous Image Generation

Kilichbek Haydarov, Aashiq Muhamed, Xiaoqian Shen, Jovana Lazarevic, Ivan Skorok hodov, Chamuditha Jayanga Galappaththige, Mohamed Elhoseiny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6316-6326

Existing GAN-based text-to-image models treat images as 2D pixel arrays. In this paper we approach the text-to-image task from a different perspective where a 2 D image is represented as an implicit neural representation (INR). We show that straightforward conditioning of the unconditional INR-based GAN method on text i nputs is not enough to achieve good performance. We propose a word-level attenti on-based weight modulation operator that controls the generation process of INR-GAN based on hypernetworks. Our experiments on benchmark datasets show that Hype rCGAN achieves competitive performance to existing pixel-based methods and retai ns the properties of continuous generative models.
****

The Neglected Tails in Vision-Language Models

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, J ames Caverlee, Shu Kong; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2024, pp. 12988-12997

Vision-language models (VLMs) excel in zero-shot recognition but their performan ce varies greatly across different visual concepts. For example although CLIP ac hieves impressive accuracy on ImageNet (60-80%) its performance drops below 10% for more than ten concepts like night snake presumably due to their limited pres ence in the pretraining data. However measuring the frequency of concepts in VLM

s' large-scale datasets is challenging. We address this by using large language models (LLMs) to count the number of pretraining texts that contain synonyms of these concepts. Our analysis confirms that popular datasets such as LAION exhibit a long-tailed concept distribution yielding biased performance in VLMs. We also find that downstream applications of VLMs including visual chatbots (e.g. GPT-4V) and text-to-image models (e.g. Stable Diffusion) often fail to recognize or generate images of rare concepts identified by our method. To mitigate the imbalanced performance of zero-shot VLMs we propose REtrieval-Augmented Learning (REAL). First instead of prompting VLMs using the original class names REAL uses their most frequent synonyms found in pretraining texts. This simple change already outperforms costly human-engineered and LLM-enriched prompts over nine benchmark datasets. Second REAL trains a linear classifier on a small yet balanced set of pretraining data retrieved using concept synonyms. REAL surpasses the previous zero-shot SOTA using 400x less storage and 10000x less training time!
****
Learning Background Prompts to Discover Implicit Knowledge for Open Vocabulary Object Detection
Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, Guanbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16678-16687
Open vocabulary object detection (OVD) aims at seeking an optimal object detector capable of recognizing objects from both base and novel categories. Recent advances leverage knowledge distillation to transfer insightful knowledge from pre-trained large-scale vision-language models to the task of object detection significantly generalizing the powerful capabilities of the detector to identify more unknown object categories. However these methods face significant challenges in background interpretation and model overfitting and thus often result in the loss of crucial background knowledge giving rise to sub-optimal inference performance of the detector. To mitigate these issues we present a novel OVD framework termed LBP to propose learning background prompts to harness explored implicit background knowledge thus enhancing the detection performance w.r.t. base and novel categories. Specifically we devise three modules: Background Category-specific Prompt Background Object Discovery and Inference Probability Rectification to empower the detector to discover represent and leverage implicit object knowledge explored from background proposals. Evaluation on two benchmark datasets OV-COCO and OV-LVIS demonstrates the superiority of our proposed method over existing state-of-the-art approaches in handling the OVD tasks.
****
HumanNeRF-SE: A Simple yet Effective Approach to Animate HumanNeRF with Diverse Poses
Caoyuan Ma, Yu-Lun Liu, Zhixiang Wang, Wu Liu, Xinchen Liu, Zheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1460-1470
We present HumanNeRF-SE a simple yet effective method that synthesizes diverse novel pose images with simple input. Previous HumanNeRF works require a large number of optimizable parameters to fit the human images. Instead we reload these approaches by combining explicit and implicit human representations to design both generalized rigid deformation and specific non-rigid deformation. Our key insight is that explicit shape can reduce the sampling points used to fit implicit representation and frozen blending weights from SMPL constructing a generalized rigid deformation can effectively avoid overfitting and improve pose generalization performance. Our architecture involving both explicit and implicit representation is simple yet effective. Experiments demonstrate our model can synthesize images under arbitrary poses with few-shot input and increase the speed of synthesizing images by 15 times through a reduction in computational complexity without using any existing acceleration modules. Compared to the state-of-the-art HumanNeRF studies HumanNeRF-SE achieves better performance with fewer learnable parameters and less training time.
****
HOLD: Category-agnostic 3D Reconstruction of Interacting Hands and Objects from

Video
Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J. Black, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 494-504
Since humans interact with diverse objects every day the holistic 3D capture of these interactions is important to understand and model human behaviour. However most existing methods for hand-object reconstruction from RGB either assume pre-scanned object templates or heavily rely on limited 3D hand-object data restricting their ability to scale and generalize to more unconstrained interaction settings. To address this we introduce HOLD -- the first category-agnostic method that reconstructs an articulated hand and an object jointly from a monocular interaction video. We develop a compositional articulated implicit model that can reconstruct disentangled 3D hands and objects from 2D images. We also further incorporate hand-object constraints to improve hand-object poses and consequently the reconstruction quality. Our method does not rely on any 3D hand-object annotations while significantly outperforming fully-supervised baselines in both in-the-lab and challenging in-the-wild settings. Moreover we qualitatively show its robustness in reconstructing from in-the-wild videos. See https://github.com/zc-alexfan/hold for code data models and updates.
****
Continual Segmentation with Disentangled Objectness Learning and Class Recognition
Yizheng Gong, Siyue Yu, Xiaoyang Wang, Jimin Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3848-3857
Most continual segmentation methods tackle the problem as a per-pixel classification task. However such a paradigm is very challenging and we find query-based segmenters with built-in objectness have inherent advantages compared with per-pixel ones as objectness has strong transfer ability and forgetting resistance. Based on these findings we propose CoMasTRe by disentangling continual segmentation into two stages: forgetting-resistant continual objectness learning and well-researched continual classification. CoMasTRe uses a two-stage segmenter learning class-agnostic mask proposals at the first stage and leaving recognition to the second stage. During continual learning a simple but effective distillation is adopted to strengthen objectness. To further mitigate the forgetting of old classes we design a multi-label class distillation strategy suited for segmentation. We assess the effectiveness of CoMasTRe on PASCAL VOC and ADE20K. Extensive experiments show that our method outperforms per-pixel and query-based methods on both datasets. Code will be available at https://github.com/jordangong/CoMasTRe.
****
Towards Accurate Post-training Quantization for Diffusion Models
Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16026-16035
In this paper we propose an accurate post-training quantization framework of diffusion models (APQ-DM) for efficient image generation. Conventional quantization frameworks learn shared quantization functions for tensor discretization regardless of the generation timesteps in diffusion models while the activation distribution differs significantly across various timesteps. Meanwhile the calibration images are acquired in random timesteps which fail to provide sufficient information for generalizable quantization function learning. Both issues cause sizable quantization errors with obvious image generation performance degradation. On the contrary we design distribution-aware quantization functions for activation discretization in different timesteps and search the optimal timesteps for informative calibration image generation so that our quantized diffusion model can reduce the discretization errors with negligible computational overhead. Specifically we partition various timestep quantization functions into different groups according to the importance weights which are optimized by differentiable search algorithms. We also extend structural risk minimization principle for informative calibration image generation to enhance the generalization ability in the deployment of quantized diffusion model. Extensive experimental results show that ou

r method outperforms the state-of-the-art post-training quantization of diffusion model by a sizable margin with similar computational cost.
****
ASAM: Boosting Segment Anything Model with Adversarial Tuning
Bo Li, Haoke Xiao, Lv Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3699-3710
In the evolving landscape of computer vision foundation models have emerged as pivotal tools exhibiting exceptional adaptability to a myriad of tasks. Among these the Segment Anything Model (SAM) by Meta AI has distinguished itself in image segmentation. However SAM like its counterparts encounters limitations in specific niche applications prompting a quest for enhancement strategies that do not compromise its inherent capabilities. This paper introduces ASAM a novel methodology that amplifies SAM's performance through adversarial tuning. We harness the potential of natural adversarial examples inspired by their successful implementation in natural language processing. By utilizing a stable diffusion model we augment a subset (1%) of the SA-1B dataset generating adversarial instances that are more representative of natural variations rather than conventional imperceptible perturbations. Our approach maintains the photorealism of adversarial examples and ensures alignment with original mask annotations thereby preserving the integrity of the segmentation task. The fine-tuned ASAM demonstrates significant improvements across a diverse range of segmentation tasks without necessitating additional data or architectural modifications. The results of our extensive evaluations confirm that ASAM establishes new benchmarks in segmentation tasks thereby contributing to the advancement of foundational models in computer vision. Our project page is in https://asam2024.github.io/.
****
UniBind: LLM-Augmented Unified and Balanced Representation Space to Bind Them All
Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26752-26762
We present UniBind a flexible and efficient approach that learns a unified representation space for seven diverse modalities-- images text audio point cloud thermal video and event data. Existing works eg. ImageBind treat the image as the central modality and build an image-centered representation space; however the space may be sub-optimal as it leads to an unbalanced representation space among all modalities. Moreover the category names are directly used to extract text embeddings for the downstream tasks making it hardly possible to represent the semantics of multi-modal data. The 'out-of-the-box' insight of our UniBind is to make the alignment center modality-agnostic and further learn a unified and balanced representation space empowered by the large language models (LLMs). UniBind is superior in its flexible application to all CLIP-style models and delivers remarkable performance boosts. To make this possible we 1) construct a knowledge base of text embeddings with the help of LLMs and multi-modal LLMs; 2) adaptively build LLM-augmented class-wise embedding center on top of the knowledge base and encoded visual embeddings; 3) align all the embeddings to the LLM-augmented embedding center via contrastive learning to achieve a unified and balanced representation space. UniBind shows strong zero-shot recognition performance gains over prior arts by an average of 6.36%. Finally we achieve new state-of-the-art performance eg. a 6.75% gain on ImageNet on the multi-modal fine-tuning setting while reducing 90% of the learnable parameters.
****
Dynamic Support Information Mining for Category-Agnostic Pose Estimation
Pengfei Ren, Yuanyuan Gao, Haifeng Sun, Qi Qi, Jingyu Wang, Jianxin Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1921-1930
Category-agnostic pose estimation (CAPE) aims to predict the pose of a query image based on few support images with pose annotations. Existing methods achieve the localization of arbitrary keypoints through similarity matching between support keypoint features and query image features. However these methods primarily focus on mining information from the query images neglecting the fact that suppor

t samples with keypoint annotations contain rich category-specific fine-grained semantic information and prior structural information. In this paper we propose a Support-based Dynamic Perception Network (SDPNet) for the robust and accurate CAPE. On the one hand SDPNet models complex dependencies between support keypoints constructing category-specific prior structure to guide the interaction of query keypoints. On the other hand SDPNet extracts fine-grained semantic information from support samples dynamically modulating the refinement process of query. Our method outperforms existing methods on MP-100 dataset by a large margin.
****
Test-Time Adaptation for Depth Completion
Hyoungseob Park, Anjali Gupta, Alex Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20519-20529
It is common to observe performance degradation when transferring models trained on some (source) datasets to target testing data due to a domain gap between them. Existing methods for bridging this gap such as domain adaptation (DA) may require the source data on which the model was trained (often not available) while others i.e. source-free DA require many passes through the testing data. We propose an online test-time adaptation method for depth completion the task of inferring a dense depth map from a single image and associated sparse depth map that closes the performance gap in a single pass. We first present a study on how the domain shift in each data modality affects model performance. Based on our observations that the sparse depth modality exhibits a much smaller covariate shift than the image we design an embedding module trained in the source domain that preserves a mapping from features encoding only sparse depth to those encoding image and sparse depth. During test time sparse depth features are projected using this map as a proxy for source domain features and are used as guidance to train a set of auxiliary parameters (i.e. adaptation layer) to align image and sparse depth features from the target test domain to that of the source domain. We evaluate our method on indoor and outdoor scenarios and show that it improves over baselines by an average of 21.1%. Code available at https://github.com/seobbro/TTA-depth-completion.
****
GOAT-Bench: A Benchmark for Multi-Modal Lifelong Navigation
Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16373-16383
The Embodied AI community has recently made significant strides in visual navigation tasks exploring targets from 3D coordinates objects language description and images. However these navigation models often handle only a single input modality as the target. With the progress achieved so far it is time to move towards universal navigation models capable of handling various goal types enabling more effective user interaction with robots. To facilitate this goal we propose GOAT-Bench a benchmark for the universal navigation task referred to as GO to AnyThing (GOAT). In this task the agent is directed to navigate to a sequence of targets specified by the category name language description or instance image in an open-vocabulary fashion. We benchmark monolithic RL and modular methods on the GOAT task analyzing their performance across modalities the role of explicit and implicit scene memories their robustness to noise in goal specifications and the impact of memory in lifelong scenarios.
****
Taming Mode Collapse in Score Distillation for Text-to-3D Generation
Peihao Wang, Dejia Xu, Zhiwen Fan, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, Vikas Chandra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9037-9047
Despite the remarkable performance of score distillation in text-to-3D generation such techniques notoriously suffer from view inconsistency issues also known as "Janus" artifact where the generated objects fake each view with multiple front faces. Although empirically effective methods have approached this problem via

score debiasing or prompt engineering a more rigorous perspective to explain an
d tackle this problem remains elusive. In this paper we reveal that the existing
score distillation-based text-to-3D generation frameworks degenerate to maximal
likelihood seeking on each view independently and thus suffer from the mode col
lapse problem manifesting as the Janus artifact in practice. To tame mode collap
se we improve score distillation by re-establishing the entropy term in the corr
esponding variational objective which is applied to the distribution of rendered
images. Maximizing the entropy encourages diversity among different views in ge
nerated 3D assets thereby mitigating the Janus problem. Based on this new object
ive we derive a new update rule for 3D score distillation dubbed Entropic Score
Distillation (ESD). We theoretically reveal that ESD can be simplified and imple
mented by just adopting the classifier-free guidance trick upon variational scor
e distillation. Although embarrassingly straightforward our extensive experiment
s demonstrate that ESD can be an effective treatment for Janus artifacts in scor
e distillation.
****
Binarized Low-light Raw Video Enhancement
Gengchen Zhang, Yulun Zhang, Xin Yuan, Ying Fu; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25753-25762
Recently deep neural networks have achieved excellent performance on low-light r
aw video enhancement. However they often come with high computational complexity
and large memory costs which hinder their applications on resource-limited devi
ces. In this paper we explore the feasibility of applying the extremely compact
binary neural network (BNN) to low-light raw video enhancement. Nevertheless the
re are two main issues with binarizing video enhancement models. One is how to f
use the temporal information to improve low-light denoising without complex modu
les. The other is how to narrow the performance gap between binary convolutions
with the full precision ones. To address the first issue we introduce a spatial-
temporal shift operation which is easy-to-binarize and effective. The temporal s
hift efficiently aggregates the features of neighbor frames and the spatial shif
t handles the misalignment caused by the large motion in videos. For the second
issue we present a distribution-aware binary convolution which captures the dist
ribution characteristics of real-valued input and incorporates them into plain b
inary convolutions to alleviate the degradation in performance. Extensive quanti
tative and qualitative experiments have shown our high-efficiency binarized low-
light raw video enhancement method can attain a promising performance. The code
is available at https://github.com/ying-fu/BRVE.
****
MorpheuS: Neural Dynamic 360deg Surface Reconstruction from Monocular RGB-D Vide
o
Hengyi Wang, Jingwen Wang, Lourdes Agapito; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20965-20976
Neural rendering has demonstrated remarkable success in dynamic scene reconstruc
tion. Thanks to the expressiveness of neural representations prior works can acc
urately capture the motion and achieve high-fidelity reconstruction of the targe
t object. Despite this real-world video scenarios often feature large unobserved
regions where neural representations struggle to achieve realistic completion.
To tackle this challenge we introduce MorpheuS a framework for dynamic 360deg su
rface reconstruction from a casually captured RGB-D video. Our approach models t
he target scene as a canonical field that encodes its geometry and appearance in
conjunction with a deformation field that warps points from the current frame t
o the canonical space. We leverage a view-dependent diffusion prior and distill
knowledge from it to achieve realistic completion of unobserved regions. Experim
ental results on various real-world and synthetic datasets show that our method
can achieve high-fidelity 360deg surface reconstruction of a deformable object f
rom a monocular RGB-D video.
****
Decoupling Static and Hierarchical Motion Perception for Referring Video Segment
ation
Shuting He, Henghui Ding; Proceedings of the IEEE/CVF Conference on Computer Vis

ion and Pattern Recognition (CVPR), 2024, pp. 13332-13341
Referring video segmentation relies on natural language expressions to identify and segment objects often emphasizing motion clues. Previous works treat a sentence as a whole and directly perform identification at the video-level mixing up static image-level cues with temporal motion cues. However image-level features cannot well comprehend motion cues in sentences and static cues are not crucial for temporal perception. In fact static cues can sometimes interfere with temporal perception by overshadowing motion cues. In this work we propose to decouple video-level referring expression understanding into static and motion perception with a specific emphasis on enhancing temporal comprehension. Firstly we introduce an expression-decoupling module to make static cues and motion cues perform their distinct role alleviating the issue of sentence embeddings overlooking motion cues. Secondly we propose a hierarchical motion perception module to capture temporal information effectively across varying timescales. Furthermore we employ contrastive learning to distinguish the motions of visually similar objects. These contributions yield state-of-the-art performance across five datasets including a remarkable 9.2% J&F improvement on the challenging MeViS dataset.
****
MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model
Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1481-1490
This paper studies the human image animation task which aims to generate a video of a certain reference identity following a particular motion sequence. Existing animation works typically employ the frame-warping technique to animate the reference image towards the target motion. Despite achieving reasonable results these approaches face challenges in maintaining temporal consistency throughout the animation due to the lack of temporal modeling and poor preservation of reference identity. In this work we introduce MagicAnimate a diffusion-based framework that aims at enhancing temporal consistency preserving reference image faithfully and improving animation fidelity. To achieve this we first develop a video diffusion model to encode temporal information. Second to maintain the appearance coherence across frames we introduce a novel appearance encoder to retain the intricate details of the reference image. Leveraging these two innovations we further employ a simple video fusion technique to encourage smooth transitions for long video animation. Empirical results demonstrate the superiority of our method over baseline approaches on two benchmarks. Notably our approach outperforms the strongest baseline by over 38% in terms of video fidelity on the challenging TikTok dancing dataset. Code and model will be made available at https://showlab.github.io/magicanimate.
****
Dense Vision Transformer Compression with Few Samples
Hanxiao Zhang, Yifan Zhou, Guo-Hua Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15825-15834
Few-shot model compression aims to compress a large model into a more compact one with only a tiny training set (even without labels). Block-level pruning has recently emerged as a leading technique in achieving high accuracy and low latency in few-shot CNN compression. But few-shot compression for Vision Transformers (ViT) remains largely unexplored which presents a new challenge. In particular the issue of sparse compression exists in traditional CNN few-shot methods which can only produce very few compressed models of different model sizes. This paper proposes a novel framework for few-shot ViT compression named DC-ViT. Instead of dropping the entire block DC-ViT selectively eliminates the attention module while retaining and reusing portions of the MLP module. DC-ViT enables dense compression which outputs numerous compressed models that densely populate the range of model complexity. DC-ViT outperforms state-of-the-art few-shot compression methods by a significant margin of 10 percentage points along with lower latency in the compression of ViT and its variants.
****
Masked AutoDecoder is Effective Multi-Task Vision Generalist

Han Qiu, Jiaxing Huang, Peng Gao, Lewei Lu, Xiaoqin Zhang, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14152-14161
Inspired by the success of general-purpose models in NLP recent studies attempt to unify different vision tasks in the same sequence format and employ autoregressive Transformers for sequence prediction. They apply uni-directional attention to capture sequential dependencies and generate task sequences recursively. However such autoregressive Transformers may not fit vision tasks well as vision task sequences usually lack the sequential dependencies typically observed in natural languages. In this work we design Masked AutoDecoder (MAD) an effective multi-task vision generalist. MAD consists of two core designs. First we develop a parallel decoding framework that introduces bi-directional attention to capture contextual dependencies comprehensively and decode vision task sequences in parallel. Second we design a masked sequence modeling approach that learns rich task contexts by masking and reconstructing task sequences. In this way MAD handles all the tasks by a single network branch and a simple cross-entropy loss with minimal task-specific designs. Extensive experiments demonstrate the great potential of MAD as a new paradigm for unifying various vision tasks. MAD achieves superior performance and inference efficiency compared to autoregressive counterparts while obtaining competitive accuracy with task-specific models. Code will be released at https://github.com/hanqiu-hq/MAD.
****
Weakly Misalignment-free Adaptive Feature Alignment for UAVs-based Multimodal Object Detection
Chen Chen, Jiahao Qi, Xingyue Liu, Kangcheng Bin, Ruigang Fu, Xikun Hu, Ping Zhong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26836-26845
Visible-infrared (RGB-IR) image fusion has shown great potentials in object detection based on unmanned aerial vehicles (UAVs). However the weakly misalignment problem between multimodal image pairs limits its performance in object detection. Most existing methods often ignore the modality gap and emphasize a strict alignment resulting in an upper bound of alignment quality and an increase of implementation costs. To address these challenges we propose a novel method named Offset-guided Adaptive Feature Alignment (OAFA) which could adaptively adjust the relative positions between multimodal features. Considering the impact of modality gap on the cross-modality spatial matching a Cross-modality Spatial Offset Modeling (CSOM) module is designed to establish a common subspace to estimate the precise feature-level offsets. Then an Offset-guided Deformable Alignment and Fusion (ODAF) module is utilized to implicitly capture optimal fusion positions for detection task rather than conducting a strict alignment. Comprehensive experiments demonstrate that our method not only achieves state-of-the-art performance in the UAVs-based object detection task but also shows strong robustness to the weakly misalignment problem.
****
From Correspondences to Pose: Non-minimal Certifiably Optimal Relative Pose without Disambiguation
Javier Tirado-Garín, Javier Civera; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 403-412
Estimating the relative camera pose from n \geq 5 correspondences between two calibrated views is a fundamental task in computer vision. This process typically involves two stages: 1) estimating the essential matrix between the views and 2) disambiguating among the four candidate relative poses that satisfy the epipolar geometry. In this paper we demonstrate a novel approach that for the first time bypasses the second stage. Specifically we show that it is possible to directly estimate the correct relative camera pose from correspondences without needing a post-processing step to enforce the cheirality constraint on the correspondences. Building on recent advances in certifiable non-minimal optimization we frame the relative pose estimation as a Quadratically Constrained Quadratic Program (QCQP). By applying the appropriate constraints we ensure the estimation of a camera pose that corresponds to a valid 3D geometry and that is globally optimal w

hen certified. We validate our method through exhaustive synthetic and real-worl
d experiments confirming the efficacy efficiency and accuracy of the proposed ap
proach. Code is available at https://github.com/javrtg/C2P.
****
Passive Snapshot Coded Aperture Dual-Pixel RGB-D Imaging
Bhargav Ghanekar, Salman Siddique Khan, Pranav Sharma, Shreyas Singh, Vivek Boom
inathan, Kaushik Mitra, Ashok Veeraraghavan; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25348-25357
Passive compact single-shot 3D sensing is useful in many application areas such
as microscopy medical imaging surgical navigation and autonomous driving where f
orm factor time and power constraints can exist. Obtaining RGB-D scene informati
on over a short imaging distance in an ultra-compact form factor and in a passiv
e snapshot manner is challenging. Dual-pixel (DP) sensors are a potential soluti
on to achieve the same. DP sensors collect light rays from two different halves
of the lens in two interleaved pixel arrays thus capturing two slightly differen
t views of the scene like a stereo camera system. However imaging with a DP sens
or implies that the defocus blur size is directly proportional to the disparity
seen between the views. This creates a trade-off between disparity estimation vs
. deblurring accuracy. To improve this trade-off effect we propose CADS (Coded A
perture Dual-Pixel Sensing) in which we use a coded aperture in the imaging lens
 along with a DP sensor. In our approach we jointly learn an optimal coded patte
rn and the reconstruction algorithm in an end-to-end optimization setting. Our r
esulting CADS imaging system demonstrates improvement of >1.5dB PSNR in all-in-f
ocus (AIF) estimates and 5-6% in depth estimation quality over naive DP sensing
for a wide range of aperture settings. Furthermore we build the proposed CADS pr
ototypes for DSLR photography settings and in an endoscope and a dermoscope form
 factor. Our novel coded dual-pixel sensing approach demonstrates accurate RGB-D
 reconstruction results in simulations and real-world experiments in a passive s
napshot and compact manner.
****
Loose Inertial Poser: Motion Capture with IMU-attached Loose-Wear Jacket
Chengxu Zuo, Yiming Wang, Lishuang Zhan, Shihui Guo, Xinyu Yi, Feng Xu, Yipeng Q
in; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2024, pp. 2209-2219
Existing wearable motion capture methods typically demand tight on-body fixation
 (often using straps) for reliable sensing limiting their application in everyda
y life. In this paper we introduce Loose Inertial Poser a novel motion capture s
olution with high wearing comfortableness by integrating four Inertial Measureme
nt Units (IMUs) into a loose-wear jacket. Specifically we address the challenge
of scarce loose-wear IMU training data by proposing a Secondary Motion AutoEncod
er (SeMo-AE) that learns to model and synthesize the effects of secondary motion
 between the skin and loose clothing on IMU data. SeMo-AE is leveraged to genera
te a diverse synthetic dataset of loose-wear IMU data to augment training for th
e pose estimation network and significantly improve its accuracy. For validation
 we collected a dataset with various subjects and 2 wearing styles (zipped and u
nzipped). Experimental results demonstrate that our approach maintains high-qual
ity real-time posture estimation even in loose-wear scenarios.
****
Instance Tracking in 3D Scenes from Egocentric Videos
Yunhan Zhao, Haoyu Ma, Shu Kong, Charless Fowlkes; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21933-219
44
Egocentric sensors such as AR/VR devices capture human-object interactions and o
ffer the potential to provide task-assistance by recalling 3D locations of objec
ts of interest in the surrounding environment. This capability requires instance
 tracking in real-world 3D scenes from egocentric videos (IT3DEgo). We explore t
his problem by first introducing a new benchmark dataset consisting of RGB and d
epth videos per-frame camera pose and instance-level annotations in both 2D came
ra and 3D world coordinates. We present an evaluation protocol which evaluates t
racking performance in 3D coordinates with two settings for enrolling instances

to track: (1) single-view online enrollment where an instance is specified on-the-fly based on the human wearer's interactions. and (2) multi-view pre-enrollment where images of an instance to be tracked are stored in memory ahead of time. To address IT3DEgo we first re-purpose methods from relevant areas e.g. single object tracking (SOT) -- running SOT methods to track instances in 2D frames and lifting them to 3D using camera pose and depth. We also present a simple method that leverages pretrained segmentation and detection models to generate proposals from RGB frames and match proposals with enrolled instance images. Our experiments show that our method (with no finetuning) significantly outperforms SOT-based approaches in the egocentric setting. We conclude by arguing that the problem of egocentric instance tracking is made easier by leveraging camera pose and using a 3D allocentric (world) coordinate representation.
****
Correlation-aware Coarse-to-fine MLPs for Deformable Medical Image Registration
Mingyuan Meng, Dagan Feng, Lei Bi, Jinman Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9645-9654
Deformable image registration is a fundamental step for medical image analysis. Recently transformers have been used for registration and outperformed Convolutional Neural Networks (CNNs). Transformers can capture long-range dependence among image features which have been shown beneficial for registration. However due to the high computation/memory loads of self-attention transformers are typically used at downsampled feature resolutions and cannot capture fine-grained long-range dependence at the full image resolution. This limits deformable registration as it necessitates precise dense correspondence between each image pixel. Multi-layer Perceptrons (MLPs) without self-attention are efficient in computation/memory usage enabling the feasibility of capturing fine-grained long-range dependence at full resolution. Nevertheless MLPs have not been extensively explored for image registration and are lacking the consideration of inductive bias crucial for medical registration tasks. In this study we propose the first correlation-aware MLP-based registration network (CorrMLP) for deformable medical image registration. Our CorrMLP introduces a correlation-aware multi-window MLP block in a novel coarse-to-fine registration architecture which captures fine-grained multi-range dependence to perform correlation-aware coarse-to-fine registration. Extensive experiments with seven public medical datasets show that our CorrMLP outperforms state-of-the-art deformable registration methods.
****
Toward Generalist Anomaly Detection via In-context Residual Learning with Few-shot Sample Prompts
Jiawen Zhu, Guansong Pang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17826-17836
This paper explores the problem of Generalist Anomaly Detection (GAD) aiming to train one single detection model that can generalize to detect anomalies in diverse datasets from different application domains without any further training on the target data. Some recent studies have shown that large pre-trained Visual-Language Models (VLMs) like CLIP have strong generalization capabilities on detecting industrial defects from various datasets but their methods rely heavily on handcrafted text prompts about defects making them difficult to generalize to anomalies in other applications e.g. medical image anomalies or semantic anomalies in natural images. In this work we propose to train a GAD model with few-shot normal images as sample prompts for AD on diverse datasets on the fly. To this end we introduce a novel approach that learns an in-context residual learning model for GAD termed InCTRL. It is trained on an auxiliary dataset to discriminate anomalies from normal samples based on a holistic evaluation of the residuals between query images and few-shot normal sample prompts. Regardless of the datasets per definition of anomaly larger residuals are expected for anomalies than normal samples thereby enabling InCTRL to generalize across different domains without further training. Comprehensive experiments on nine AD datasets are performed to establish a GAD benchmark that encapsulate the detection of industrial defect anomalies medical anomalies and semantic anomalies in both one-vs-all and multi-class setting on which InCTRL is the best performer and significantly outperform

s state-of-the-art competing methods. Code is available at https://github.com/ma la-lab/InCTRL.
****
Fourier-basis Functions to Bridge Augmentation Gap: Rethinking Frequency Augmentation in Image Classification
Puru Vaish, Shunxin Wang, Nicola Strisciuglio; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17763-17772
Computer vision models normally witness degraded performance when deployed in real-world scenarios due to unexpected changes in inputs that were not accounted for during training. Data augmentation is commonly used to address this issue as it aims to increase data variety and reduce the distribution gap between training and test data. However common visual augmentations might not guarantee extensive robustness of computer vision models. In this paper we propose Auxiliary Fourier-basis Augmentation (AFA) a complementary technique targeting augmentation in the frequency domain and filling the robustness gap left by visual augmentations. We demonstrate the utility of augmentation via Fourier-basis additive noise in a straightforward and efficient adversarial setting. Our results show that AFA benefits the robustness of models against common corruptions OOD generalization and consistency of performance of models against increasing perturbations with negligible deficit to the standard performance of models. It can be seamlessly integrated with other augmentation techniques to further boost performance. Codes and models are available at \href https://github.com/nis-research/afa-augment https://github.com/nis-research/afa-augment .
****
Learning to Transform Dynamically for Better Adversarial Transferability
Rongyi Zhu, Zeliang Zhang, Susan Liang, Zhuo Liu, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24273-24283
Adversarial examples crafted by adding perturbations imperceptible to humans can deceive neural networks. Recent studies identify the adversarial transferability across various models i.e. the cross-model attack ability of adversarial samples. To enhance such adversarial transferability existing input transformation-based methods diversify input data with transformation augmentation. However their effectiveness is limited by the finite number of available transformations. In our study we introduce a novel approach named Learning to Transform (L2T). L2T increases the diversity of transformed images by selecting the optimal combination of operations from a pool of candidates consequently improving adversarial transferability. We conceptualize the selection of optimal transformation combinations as a trajectory optimization problem and employ a reinforcement learning strategy to effectively solve the problem. Comprehensive experiments on the ImageNet dataset as well as practical tests with Google Vision and GPT-4V reveal that L2T surpasses current methodologies in enhancing adversarial transferability thereby confirming its effectiveness and practical significance.
****
PlatoNeRF: 3D Reconstruction in Plato's Cave via Single-View Two-Bounce Lidar
Tzofi Klinghoffer, Xiaoyu Xiang, Siddharth Somasundaram, Yuchen Fan, Christian Richardt, Ramesh Raskar, Rakesh Ranjan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14565-14574
3D reconstruction from a single-view is challenging because of the ambiguity from monocular cues and lack of information about occluded regions. Neural radiance fields (NeRF) while popular for view synthesis and 3D reconstruction are typically reliant on multi-view images. Existing methods for single-view 3D reconstruction with NeRF rely on either data priors to hallucinate views of occluded regions which may not be physically accurate or shadows observed by RGB cameras which are difficult to detect in ambient light and low albedo backgrounds. We propose using time-of-flight data captured by a single-photon avalanche diode to overcome these limitations. Our method models two-bounce optical paths with NeRF using lidar transient data for supervision. By leveraging the advantages of both NeRF and two-bounce light measured by lidar we demonstrate that we can reconstruct visible and occluded geometry without data priors or reliance on controlled ambie

nt lighting or scene albedo. In addition we demonstrate improved generalization under practical constraints on sensor spatial- and temporal-resolution. We believe our method is a promising direction as single-photon lidars become ubiquitous on consumer devices such as phones tablets and headsets.
****
PanoContext-Former: Panoramic Total Scene Understanding with a Transformer
Yuan Dong, Chuan Fang, Liefeng Bo, Zilong Dong, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28087-28097
Panoramic images enable deeper understanding and more holistic perception of 360 surrounding environment which can naturally encode enriched scene context information compared to standard perspective image. Previous work has made lots of effort to solve the scene understanding task in a hybrid solution based on 2D-3D geometric reasoning thus each sub-task is processed separately and few correlations are explored in this procedure. In this paper we propose a fully 3D method for holistic indoor scene understanding which recovers the objects' shapes oriented bounding boxes and the 3D room layout simultaneously from a single panorama. To maximize the exploration of the rich context information we design a transformer-based context module to predict the representation and relationship among each component of the scene. In addition we introduce a new dataset for scene understanding including photo-realistic panoramas high-fidelity depth images accurately annotated room layouts oriented object bounding boxes and shapes. Experiments on the synthetic and new datasets demonstrate that our method outperforms previous panoramic scene understanding methods in terms of both layout estimation and 3D object detection.
****
Training-Free Pretrained Model Merging
Zhengqi Xu, Ke Yuan, Huiqiong Wang, Yong Wang, Mingli Song, Jie Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5915-5925
Recently model merging techniques have surfaced as a solution to combine multiple single-talent models into a single multi-talent model. However previous endeavors in this field have either necessitated additional training or fine-tuning processes or require that the models possess the same pre-trained initialization. In this work we identify a common drawback in prior works w.r.t. the inconsistency of unit similarity in the weight space and the activation space. To address this inconsistency we propose an innovative model merging framework coined as merging under dual-space constraints (MuDSC). Specifically instead of solely maximizing the objective of a single space we advocate for the exploration of permutation matrices situated in a region with a unified high similarity in the dual space achieved through the linear combination of activation and weight similarity matrices. In order to enhance usability we have also incorporated adaptations for group structure including Multi-Head Attention and Group Normalization. Comprehensive experimental comparisons demonstrate that MuDSC can significantly boost the performance of merged models with various task combinations and architectures. Furthermore the visualization of the merged model within the multi-task loss landscape reveals that MuDSC enables the merged model to reside in the overlapping segment featuring a unified lower loss for each task. Our code is publicly available at https://github.com/zju-vipa/training_free_model_merging.
****
NC-SDF: Enhancing Indoor Scene Reconstruction Using Neural SDFs with View-Dependent Normal Compensation
Ziyi Chen, Xiaolong Wu, Yu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5155-5165
State-of-the-art neural implicit surface representations have achieved impressive results in indoor scene reconstruction by incorporating monocular geometric priors as additional supervision. However we have observed that multi-view inconsistency between such priors poses a challenge for high-quality reconstructions. In response we present NC-SDF a neural signed distance field (SDF) 3D reconstruction framework with view-dependent normal compensation (NC). Specifically we inte

grate view-dependent biases in monocular normal priors into the neural implicit representation of the scene. By adaptively learning and correcting the biases ou r NC-SDF effectively mitigates the adverse impact of inconsistent supervision en hancing both the global consistency and local details in the reconstructions. To further refine the details we introduce an informative pixel sampling strategy to pay more attention to intricate geometry with higher information content. Add itionally we design a hybrid geometry modeling approach to improve the neural im plicit representation. Experiments on synthetic and real-world datasets demonstr ate that NC-SDF outperforms existing approaches in terms of reconstruction quali ty.
****
An Interactive Navigation Method with Effect-oriented Affordance
Xiaohan Wang, Yuehu Liu, Xinhang Song, Yuyi Liu, Sixian Zhang, Shuqiang Jiang; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16446-16456
Visual navigation is to let the agent reach the target according to the continuo us visual input. In most previous works visual navigation is usually assumed to be done in a static and ideal environment: the target is always reachable with n o need to alter the environment. However the "messy" environments are more gener al and practical in our daily lives where the agent may get blocked by obstacles . Thus Interactive Navigation (InterNav) is introduced to navigate to the object s in more realistic "messy" environments according to the object interaction. Pr ior work on InterNav learns short-term interaction through extensive trials with reinforcement learning. However interaction does not guarantee efficient naviga tion that is planning obstacle interactions that make shorter paths and consume less effort is also crucial. In this paper we introduce an effect-oriented affor dance map to enable long-term interactive navigation extending the existing map-based navigation framework to the domain of dynamic environment. We train a set of affordance functions predicting available interactions and the time cost of r emoving obstacles which informatively support an interactive modular system to a ddress interaction and long-term planning. Experiments on the ProcTHOR simulator demonstrate the capability of our affordance-driven system in long-term navigat ion in complex dynamic environments.
****
Person in Place: Generating Associative Skeleton-Guidance Maps for Human-Object Interaction Image Editing
ChangHee Yang, ChanHee Kang, Kyeongbo Kong, Hanni Oh, Suk-Ju Kang; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 8164-8175
Recently there were remarkable advances in image editing tasks in various ways. Nevertheless existing image editing models are not designed for Human-Object Int eraction (HOI) image editing. One of these approaches (e.g. ControlNet) employs the skeleton guidance to offer precise representations of human showing better r esults in HOI image editing. However using conventional methods manually creatin g HOI skeleton guidance is necessary. This paper proposes the object interactive diffuser with associative attention that considers both the interaction with ob jects and the joint graph structure automating the generation of HOI skeleton gu idance. Additionally we propose the HOI loss with novel scaling parameter demons trating its effectiveness in generating skeletons that interact better. To evalu ate generated object-interactive skeletons we propose two metrics top-N accuracy and skeleton probabilistic distance. Our framework integrates object interactiv e diffuser that generates object-interactive skeletons with previous methods dem onstrating the outstanding results in HOI image editing. Finally we present pote ntials of our framework beyond HOI image editing as applications to human-to-hum an interaction skeleton editing and 3D mesh optimization. The code is available at https://github.com/YangChangHee/CVPR2024_Person-In-Place_RELEASE
****
PREGO: Online Mistake Detection in PRocedural EGOcentric Videos
Alessandro Flaborea, Guido Maria D'Amely di Melendugno, Leonardo Plini, Luca Sco fano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, Fabio Gala

sso; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2024, pp. 18483-18492
Promptly identifying procedural errors from egocentric videos in an online setti ng is highly challenging and valuable for detecting mistakes as soon as they hap pen. This capability has a wide range of applications across various fields such as manufacturing and healthcare. The nature of procedural mistakes is open-set since novel types of failures might occur which calls for one-class classifiers trained on correctly executed procedures. However no technique can currently det ect open-set procedural mistakes online. We propose PREGO the first online one-c lass classification model for mistake detection in PRocedural EGOcentric videos. PREGO is based on an online action recognition component to model the current a ction and a symbolic reasoning module to predict the next actions. Mistake detec tion is performed by comparing the recognized current action with the expected f uture one. We evaluate PREGO on two procedural egocentric video datasets Assembl y101 and Epic-tent which we adapt for online benchmarking of procedural mistake detection to establish suitable benchmarks thus defining the Assembly101-O and E pic-tent-O datasets respectively. The code is available at https://github.com/al eflabo/PREGO
****
ChatPose: Chatting about 3D Human Pose
Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2024, pp. 2093-2103
We introduce ChatPose a framework employing Large Language Models (LLMs) to unde rstand and reason about 3D human poses from images or textual descriptions. Our work is motivated by the human ability to intuitively understand postures from a single image or a brief description a process that intertwines image interpreta tion world knowledge and an understanding of body language. Traditional human po se estimation and generation methods often operate in isolation lacking semantic understanding and reasoning abilities. ChatPose addresses these limitations by embedding SMPL poses as distinct signal tokens within a multimodal LLM enabling the direct generation of 3D body poses from both textual and visual inputs. Leve raging the powerful capabilities of multimodal LLMs ChatPose unifies classical 3 D human pose and generation tasks while offering user interactions. Additionally ChatPose empowers LLMs to apply their extensive world knowledge in reasoning ab out human poses leading to two advanced tasks: speculative pose generation and r easoning about pose estimation. These tasks involve reasoning about humans to ge nerate 3D poses from subtle text queries possibly accompanied by images. We esta blish benchmarks for these tasks moving beyond traditional 3D pose generation an d estimation methods. Our results show that ChatPose out-performs existing multi modal LLMs and task-specific methods on these newly proposed tasks. Furthermore ChatPose's ability to understand and generate 3D human poses based on complex re asoning opens new directions in human pose analysis. Code and data are available for research at https://yfeng95.github.io/ChatPose.
****
Prompt3D: Random Prompt Assisted Weakly-Supervised 3D Object Detection
Xiaohong Zhang, Huisheng Ye, Jingwen Li, Qinyu Tang, Yuanqi Li, Yanwen Guo, Jie Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2024, pp. 28046-28055
The prohibitive cost of annotations for fully supervised 3D indoor object detect ion limits its practicality. In this work we propose Random Prompt Assisted Weak ly-supervised 3D Object Detection termed as Prompt3D a weakly-supervised approac h that leverages position-level labels to overcome this challenge. Explicitly ou r method focuses on enhancing labeling using synthetic scenes crafted from 3D sh apes generated via random prompts. First a Synthetic Scene Generation (SSG) modu le is introduced to assemble synthetic scenes with a curated collection of 3D sh apes created via random prompts for each category. These scenes are enriched wit h automatically generated point-level annotations providing a robust supervisory framework for training the detection algorithm. To enhance the transfer of know ledge from virtual to real datasets we then introduce a Prototypical Proposal Fe

ature Alignment (PPFA) module. This module effectively alleviates the domain gap by directly minimizing the distance between feature prototypes of the same class proposals across two domains. Compared with sota BR our method improves by 5.4% and 8.7% on mAP with VoteNet and GroupFree3D serving as detectors respectively demonstrating the effectiveness of our proposed method. Code is available at: https://github.com/huishengye/prompt3d.
****
Logit Standardization in Knowledge Distillation
Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, Xiaochun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15731-15740
Knowledge distillation involves transferring soft labels from a teacher to a student using a shared temperature-based softmax function. However the assumption of a shared temperature between teacher and student implies a mandatory exact match between their logits in terms of logit range and variance. This side-effect limits the performance of student considering the capacity discrepancy between them and the finding that the innate logit relations of teacher are sufficient for student to learn. To address this issue we propose setting the temperature as the weighted standard deviation of logit and performing a plug-and-play Z-score pre-process of logit standardization before applying softmax and Kullback-Leibler divergence. Our pre-process enables student to focus on essential logit relations from teacher rather than requiring a magnitude match and can improve the performance of existing logit-based distillation methods. We also show a typical case where the conventional setting of sharing temperature between teacher and student cannot reliably yield the authentic distillation evaluation; nonetheless this challenge is successfully alleviated by our Z-score. We extensively evaluate our method for various student and teacher models on CIFAR-100 and ImageNet showing its significant superiority. The vanilla knowledge distillation powered by our pre-process can achieve favorable performance against state-of-the-art methods and other distillation variants can obtain considerable gain with the assistance of our pre-process. The codes pre-trained models and logs are released on Github.
****
Fine-grained Prototypical Voting with Heterogeneous Mixup for Semi-supervised 2D-3D Cross-modal Retrieval
Fan Zhang, Xian-Sheng Hua, Chong Chen, Xiao Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17016-17026
This paper studies the problem of semi-supervised 2D-3D retrieval which aims to align both labeled and unlabeled 2D and 3D data into the same embedding space. The problem is challenging due to the complicated heterogeneous relationships between 2D and 3D data. Moreover label scarcity in real-world applications hinders from generating discriminative representations. In this paper we propose a semi-supervised approach named Fine-grained Prototypcical Voting with Heterogeneous Mixup (FIVE) which maps both 2D and 3D data into a common embedding space for cross-modal retrieval. Specifically we generate fine-grained prototypes to model inter-class variation for both 2D and 3D data. Then considering each unlabeled sample as a query we retrieve relevant prototypes to vote for reliable and robust pseudo-labels which serve as guidance for discriminative learning under label scarcity. Furthermore to bridge the semantic gap between two modalities we mix cross-modal pairs with similar semantics in the embedding space and then perform similarity learning for cross-modal discrepancy reduction in a soft manner. The whole FIVE is optimized with the consideration of sharpness to mitigate the impact of potential label noise. Extensive experiments on benchmark datasets validate the superiority of FIVE compared with a range of baselines in different settings. On average FIVE outperforms the second-best approach by 4.74% on 3D MNIST 12.94% on ModelNet10 and 22.10% on ModelNet40.
****
Leak and Learn: An Attacker's Cookbook to Train Using Leaked Data from Federated Learning
Joshua C. Zhao, Ahaan Dabholkar, Atul Sharma, Saurabh Bagchi; Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12247-12256
Federated learning is a decentralized learning paradigm introduced to preserve privacy of client data. Despite this prior work has shown that an attacker at the server can still reconstruct the private training data using only the client updates. These attacks are known as data reconstruction attacks and fall into two major categories: gradient inversion (GI) and linear layer leakage attacks (LLL). However despite demonstrating the effectiveness of these attacks in breaching privacy prior work has not investigated the usefulness of the reconstructed data for downstream tasks. In this work we explore data reconstruction attacks through the lens of training and improving models with leaked data. We demonstrate the effectiveness of both GI and LLL attacks in maliciously training models using the leaked data more accurately than a benign federated learning strategy. Counter-intuitively this bump in training quality can occur despite limited reconstruction quality or a small total number of leaked images. Finally we show the limitations of these attacks for downstream training individually for GI attacks and for LLL attacks.
****
OCAI: Improving Optical Flow Estimation by Occlusion and Consistency Aware Interpolation
Jisoo Jeong, Hong Cai, Risheek Garrepalli, Jamie Menjay Lin, Munawar Hayat, Fatih Porikli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19352-19362
The scarcity of ground-truth labels poses one major challenge in developing optical flow estimation models that are both generalizable and robust. While current methods rely on data augmentation they have yet to fully exploit the rich information available in labeled video sequences. We propose OCAI a method that supports robust frame interpolation by generating intermediate video frames alongside optical flows in between. Utilizing a forward warping approach OCAI employs occlusion awareness to resolve ambiguities in pixel values and fills in missing values by leveraging the forward-backward consistency of optical flows. Additionally we introduce a teacher-student style semi-supervised learning method on top of the interpolated frames. Using a pair of unlabeled frames and the teacher model's predicted optical flow we generate interpolated frames and flows to train a student model. The teacher's weights are maintained using Exponential Moving Averaging of the student. Our evaluations demonstrate perceptually superior interpolation quality and enhanced optical flow accuracy on established benchmarks such as Sintel and KITTI.
****
Distilling ODE Solvers of Diffusion Models into Smaller Steps
Sanghwan Kim, Hao Tang, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9410-9419
Abstract Diffusion models have recently gained prominence as a novel category of generative models. Despite their success these models face a notable drawback in terms of slow sampling speeds requiring a high number of function evaluations (NFE) in the order of hundreds or thousands. In response both learning-free and learning-based sampling strategies have been explored to expedite the sampling process. Learning-free sampling employs various ordinary differential equation (ODE) solvers based on the formulation of diffusion ODEs. However it encounters challenges in faithfully tracking the true sampling trajectory particularly for small NFE. Conversely learning-based sampling methods such as knowledge distillation demand extensive additional training limiting their practical applicability. To overcome these limitations we introduce Distilled-ODE solvers (D-ODE solvers) a straightforward distillation approach grounded in ODE solver formulations. Our method seamlessly integrates the strengths of both learning-free and learning-based sampling. D-ODE solvers are constructed by introducing a single parameter adjustment to existing ODE solvers. Furthermore we optimize D-ODE solvers with smaller steps using knowledge distillation from ODE solvers with larger steps across a batch of samples. Comprehensive experiments demonstrate the superior performance of D-ODE solvers compared to existing ODE solvers including DDIM PNDM DPM

-Solver DEIS and EDM particularly in scenarios with fewer NFE. Notably our metho
d incurs negligible computational overhead compared to previous distillation tec
hniques facilitating straightforward and rapid integration with existing sampler
s. Qualitative analysis reveals that D-ODE solvers not only enhance image qualit
y but also faithfully follow the target ODE trajectory.
****
Navigating Beyond Dropout: An Intriguing Solution towards Generalizable Image Su
per Resolution
Hongjun Wang, Jiyuan Chen, Yinqiang Zheng, Tieyong Zeng; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 255
32-25543
Deep learning has led to a dramatic leap on Single Image Super-Resolution (SISR)
 performances in recent years. While most existing work assumes a simple and fix
ed degradation model (e.g. bicubic downsampling) the research of Blind SR seeks
to improve model generalization ability with unknown degradation. Recently Kong
et al. pioneer the investigation of a more suitable training strategy for Blind
SR using Dropout. Although such method indeed brings substantial generalization
improvements via mitigating overfitting we argue that Dropout simultaneously int
roduces undesirable side-effect that compromises model's capacity to faithfully
reconstruct fine details. We show both the theoretical and experimental analyses
 in our paper and furthermore we present another easy yet effective training str
ategy that enhances the generalization ability of the model by simply modulating
 its first and second-order features statistics. Experimental results have shown
 that our method could serve as a model-agnostic regularization and outperforms
Dropout on seven benchmark datasets including both synthetic and real-world scen
arios.
****
Doodle Your 3D: From Abstract Freehand Sketches to Precise 3D Shapes
Hmrishav Bandyopadhyay, Subhadeep Koley, Ayan Das, Ayan Kumar Bhunia, Aneeshan S
ain, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9795-980
5
In this paper we democratise 3D content creation enabling precise generation of
3D shapes from abstract sketches while overcoming limitations tied to drawing sk
ills. We introduce a novel part-level modelling and alignment framework that fac
ilitates abstraction modelling and cross-modal correspondence. Leveraging the sa
me part-level decoder our approach seamlessly extends to sketch modelling by est
ablishing correspondence between CLIPasso edgemaps and projected 3D part regions
 eliminating the need for a dataset pairing human sketches and 3D shapes. Additi
onally our method introduces a seamless in-position editing process as a byprodu
ct of cross-modal part-aligned modelling. Operating in a low-dimensional implici
t space our approach significantly reduces computational demands and processing
time.
****
LightIt: Illumination Modeling and Control for Diffusion Models
Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Nießner, Yannick Hold-G
eoffroy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2024, pp. 9359-9369
We introduce LightIt a method for explicit illumination control for image genera
tion. Recent generative methods lack lighting control which is crucial to numero
us artistic aspects of image generation such as setting the overall mood or cine
matic appearance. To overcome these limitations we propose to condition the gene
ration on shading and normal maps. We model the lighting with single bounce shad
ing which includes cast shadows. We first train a shading estimation module to g
enerate a dataset of real-world images and shading pairs. Then we train a contro
l network using the estimated shading and normals as input. Our method demonstra
tes high-quality image generation and lighting control in numerous scenes. Addit
ionally we use our generated dataset to train an identity-preserving relighting
model conditioned on an image and a target shading. Our method is the first that
 enables the generation of images with controllable consistent lighting and perf

orms on par with specialized relighting state-of-the-art methods.
****
Single View Refractive Index Tomography with Neural Fields
Brandon Zhao, Aviad Levis, Liam Connor, Pratul P. Srinivasan, Katherine L. Bouman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25358-25367
Refractive Index Tomography is the inverse problem of reconstructing the continuously-varying 3D refractive index in a scene using 2D projected image measurements. Although a purely refractive field is not directly visible it bends light rays as they travel through space thus providing a signal for reconstruction. The effects of such fields appear in many scientific computer vision settings ranging from refraction due to transparent cells in microscopy to the lensing of distant galaxies caused by dark matter in astrophysics. Reconstructing these fields is particularly difficult due to the complex nonlinear effects of the refractive field on observed images. Furthermore while standard 3D reconstruction and tomography settings typically have access to observations of the scene from many viewpoints many refractive index tomography problem settings only have access to images observed from a single viewpoint. We introduce a method that leverages prior knowledge of light sources scattered throughout the refractive medium to help disambiguate the single-view refractive index tomography problem. We differentiably trace curved rays through a neural field representation of the refractive field and optimize its parameters to best reproduce the observed image. We demonstrate the efficacy of our approach by reconstructing simulated refractive fields analyze the effects of light source distribution on the recovered field and test our method on a simulated dark matter mapping problem where we successfully recover the 3D refractive field caused by a realistic dark matter distribution.
****
Neural Lineage
Runpeng Yu, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4797-4807
Given a well-behaved neural network is possible to identify its parent based on which it was tuned? In this paper we introduce a novel task known as neural lineage detection aiming at discovering lineage relationships between parent and child models. Specifically from a set of parent models neural lineage detection predicts which parent model a child model has been fine-tuned from. We propose two approaches to address this task. (1) For practical convenience we introduce a learning-free approach which integrates an approximation of the finetuning process into the neural network representation similarity metrics leading to a similarity-based lineage detection scheme. (2) For the pursuit of accuracy we introduce a learning-based lineage detector comprising encoders and a transformer detector. Through experimentation we have validated that our proposed learning-free and learning-based methods outperform the baseline in various learning settings and are adaptable to a variety of visual models. Moreover they also exhibit the ability to trace cross-generational lineage identifying not only parent models but also their ancestors.
****
Visual Layout Composer: Image-Vector Dual Diffusion Model for Design Layout Generation
Mohammad Amin Shabani, Zhaowen Wang, Difan Liu, Nanxuan Zhao, Jimei Yang, Yasutaka Furukawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9222-9231
This paper proposes an image-vector dual diffusion model for generative layout design. Distinct from prior efforts that mostly ignore element-level visual information our approach integrates the power of a pre-trained large image diffusion model to guide layout composition in a vector diffusion model by providing enhanced salient region understanding and high-level inter-element relationship reasoning. Our proposed model simultaneously operates in two domains: it generates the overall design appearance in the image domain while optimizing the size and position of each design element in the vector domain. The proposed method achieves the state-of-the-art results on several datasets and enables new layout design

applications.
****
FC-GNN: Recovering Reliable and Accurate Correspondences from Interferences
Haobo Xu, Jun Zhou, Hua Yang, Renjie Pan, Cunyan Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25213-25222
Finding correspondences between images is essential for many computer vision tasks and sparse matching pipelines have been popular for decades. However matching noise within and between images along with inconsistent keypoint detection frequently degrades the matching performance. We review these problems and thus propose: 1) a novel and unified Filtering and Calibrating (FC) approach that jointly rejects outliers and optimizes inliers and 2) leveraging both the matching context and the underlying image texture to remove matching uncertainties. Under the guidance of the above innovations we construct Filtering and Calibrating Graph Neural Network (FC-GNN) which follows the FC approach to recover reliable and accurate correspondences from various interferences. FC-GNN conducts an effectively combined inference of contextual and local information through careful embedding and multiple information aggregations predicting confidence scores and calibration offsets for the input correspondences to jointly filter out outliers and improve pixel-level matching accuracy. Moreover we exploit the local coherence of matches to perform inference on local graphs thereby reducing computational complexity. Overall FC-GNN operates at lightning speed and can greatly boost the performance of diverse matching pipelines across various tasks showcasing the immense potential of such approaches to become standard and pivotal components of image matching. Code is avaiable at https://github.com/xuy123456/fcgnn.
****
Turb-Seg-Res: A Segment-then-Restore Pipeline for Dynamic Videos with Atmospheric Turbulence
Ripon Kumar Saha, Dehao Qin, Nianyi Li, Jinwei Ye, Suren Jayasuriya; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25286-25296
Tackling image degradation due to atmospheric turbulence particularly in dynamic environments remains a challenge for long-range imaging systems. Existing techniques have been primarily designed for static scenes or scenes with small motion. This paper presents the first segment-then-restore pipeline for restoring the videos of dynamic scenes in turbulent environments. We leverage mean optical flow with an unsupervised motion segmentation method to separate dynamic and static scene components prior to restoration. After camera shake compensation and segmentation we introduce foreground/background enhancement leveraging the statistics of turbulence strength and a transformer model trained on a novel noise-based procedural turbulence generator for fast dataset augmentation. Benchmarked against existing restoration methods our approach restores most of the geometric distortion and enhances the sharpness of videos. We make our code simulator and data publicly available to advance the field of video restoration from turbulence: riponcs.github.io/TurbSegRes
****
Real-time Acquisition and Reconstruction of Dynamic Volumes with Neural Structured Illumination
Yixin Zeng, Zoubin Bi, Mingrui Yin, Xiang Feng, Kun Zhou, Hongzhi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20186-20195
We propose a novel framework for real-time acquisition and reconstruction of temporally-varying 3D phenomena with high quality. The core of our framework is a deep neural network with an encoder that directly maps to the structured illumination during acquisition a decoder that predicts a 1D density distribution from single-pixel measurements under the optimized lighting and an aggregation module that combines the predicted densities for each camera into a single volume. It enables the automatic and joint optimization of physical acquisition and computational reconstruction and is flexible to adapt to different hardware configurations. The effectiveness of our framework is demonstrated on a lightweight setup wi

th an off-the-shelf projector and one or multiple cameras achieving a performance of 40 volumes per second at a spatial resolution of 128^3. We compare favorably with state-of-the-art techniques in real and synthetic experiments and evaluate the impact of various factors over our pipeline.
****

3D Multi-frame Fusion for Video Stabilization
Zhan Peng, Xinyi Ye, Weiyue Zhao, Tianqi Liu, Huiqiang Sun, Baopu Li, Zhiguo Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7507-7516
In this paper we present RStab a novel framework for video stabilization that integrates 3D multi-frame fusion through volume rendering. Departing from conventional methods we introduce a 3D multi-frame perspective to generate stabilized images addressing the challenge of full-frame generation while preserving structure. The core of our RStab framework lies in Stabilized Rendering (SR) a volume rendering module fusing multi-frame information in 3D space. Specifically SR involves warping features and colors from multiple frames by projection fusing them into descriptors to render the stabilized image. However the precision of warped information depends on the projection accuracy a factor significantly influenced by dynamic regions. In response we introduce the Adaptive Ray Range (ARR) module to integrate depth priors adaptively defining the sampling range for the projection process. Additionally we propose Color Correction (CC) assisting geometric constraints with optical flow for accurate color aggregation. Thanks to the three modules our RStab demonstrates superior performance compared with previous stabilizers in the field of view (FOV) image quality and video stability across various datasets.
****

Local-consistent Transformation Learning for Rotation-invariant Point Cloud Analysis
Yiyang Chen, Lunhao Duan, Shanshan Zhao, Changxing Ding, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5418-5427
Rotation invariance is an important requirement for point shape analysis. To achieve this current state-of-the-art methods attempt to construct the local rotation-invariant representation through learning or defining the local reference frame (LRF). Although efficient these LRF-based methods suffer from perturbation of local geometric relations resulting in suboptimal local rotation invariance. To alleviate this issue we propose a Local-consistent Transformation (LocoTrans) learning strategy. Specifically we first construct the local-consistent reference frame (LCRF) by considering the symmetry of the two axes in LRF. In comparison with previous LRFs our LCRF is able to preserve local geometric relationships better through performing local-consistent transformation. However as the consistency only exists in local regions the relative pose information is still lost in the intermediate layers of the network. We mitigate such a relative pose issue by developing a relative pose recovery (RPR) module. RPR aims to restore the relative pose between adjacent transformed patches. Equipped with LCRF and RPR our LocoTrans is capable of learning local-consistent transformation and preserving local geometry which benefits rotation invariance learning. Competitive performance under arbitrary rotations on both shape classification and part segmentation tasks and ablations can demonstrate the effectiveness of our method. Code will be available publicly at https://github.com/wdttt/LocoTrans.
****

Tailored Visions: Enhancing Text-to-Image Generation with Personalized Prompt Rewriting
Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, Zhenzhong Lan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7727-7736
Despite significant progress in the field it is still challenging to create personalized visual representations that align closely with the desires and preferences of individual users. This process requires users to articulate their ideas in words that are both comprehensible to the models and accurately capture their

vision posing difficulties for many users. In this paper we tackle this challeng
e by leveraging historical user interactions with the system to enhance user pro
mpts. We propose a novel approach that involves rewriting user prompts based on
a newly collected large-scale text-to-image dataset with over 300k prompts from
3115 users. Our rewriting model enhances the expressiveness and alignment of use
r prompts with their intended visual outputs. Experimental results demonstrate t
he superiority of our methods over baseline approaches as evidenced in our new o
ffline evaluation method and online tests. Our code and dataset are available at
 https://github.com/zzjchen/Tailored-Visions
****
Efficient Deformable ConvNets: Rethinking Dynamic and Sparse Operator for Vision
 Applications
Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai W
ang, Tong Lu, Hongsheng Li, Yu Qiao, Lewei Lu, Jie Zhou, Jifeng Dai; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 5652-5661
We introduce Deformable Convolution v4 (DCNv4) a highly efficient and effective
operator designed for a broad spectrum of vision applications. DCNv4 addresses t
he limitations of its predecessor DCNv3 with two key enhancements: 1. removing s
oftmax normalization in spatial aggregation to enhance its dynamic property and
expressive power and 2. optimizing memory access to minimize redundant operation
s for speedup. These improvements result in a significantly faster convergence c
ompared to DCNv3 and a substantial increase in processing speed with DCNv4 achie
ving more than three times the forward speed. DCNv4 demonstrates exceptional per
formance across various tasks including image classification instance and semant
ic segmentation and notably image generation. When integrated into generative mo
dels like U-Net in the latent diffusion model DCNv4 outperforms its baseline und
erscoring its possibility to enhance generative models. In practical application
s replacing DCNv3 with DCNv4 in the InternImage model to create FlashInternImage
 results in up to 80% speed increase and further performance improvement without
 further modifications. The advancements in speed and efficiency of DCNv4 combin
ed with its robust performance across diverse vision tasks show its potential as
 a foundational building block for future vision models.
****
CoDe: An Explicit Content Decoupling Framework for Image Restoration
Enxuan Gu, Hongwei Ge, Yong Guo; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2024, pp. 2920-2930
The performance of image restoration (IR) is highly dependent on the reconstruct
ion quality of diverse contents with varying complexity. However most IR approac
hes model the mapping between various complexity contents of inputs and outputs
through the repeated feature calculation propagation mechanism in a unified pipe
line which leads to unsatisfactory results. To address this issue we propose an
explicit Content Decoupling framework for IR dubbed CoDe to end-to-end model the
 restoration process by utilizing decoupled content components in a divide-and-c
onquer-like architecture. Specifically a Content Decoupling Module is first desi
gned to decouple content components of inputs and outputs according to the frequ
ency spectra adaptively generated from the transform domain. In addition in orde
r to harness the divide-and-conquer strategy for reconstructing decoupled conten
t components we propose an IR Network Container. It contains an optimized versio
n which is a streamlining of an arbitrary IR network comprising the cascaded mod
ulated subnets and a Reconstruction Layers Pool. Finally a Content Consistency L
oss is designed from the transform domain perspective to supervise the restorati
on process of each content component and further guide the feature fusion proces
s. Extensive experiments on several IR tasks such as image super-resolution imag
e denoising and image blurring covering both real and synthetic settings demonst
rate that the proposed paradigm can effectively take the performance of the orig
inal network to a new state-of-the-art level in multiple benchmark datasets (e.g
. 0.34dB@Set5 x4 over DAT).
****
XFibrosis: Explicit Vessel-Fiber Modeling for Fibrosis Staging from Liver Pathol

ogy Images
Chong Yin, Siqi Liu, Fei Lyu, Jiahao Lu, Sune Darkner, Vincent Wai-Sun Wong, Pon
g C. Yuen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2024, pp. 11282-11291
The increasing prevalence of non-alcoholic fatty liver disease (NAFLD) has cause
d public concern in recent years. The high prevalence and risk of severe complic
ations make monitoring NAFLD progression a public health priority. Fibrosis stag
ing from liver biopsy images plays a key role in demonstrating the histological
progression of NAFLD. Fibrosis mainly involves the deposition of fibers around v
essels. Current deep learning-based fibrosis staging methods learn spatial relat
ionships between tissue patches but do not explicitly consider the relationships
 between vessels and fibers leading to limited performance and poor interpretabi
lity. In this paper we propose an eXplicit vessel-fiber modeling method for Fibr
osis staging from liver biopsy images namely XFibrosis. Specifically we transfor
m vessels and fibers into graph-structured representations where their micro-str
uctures are depicted by vessel-induced primal graphs and fiber-induced dual grap
hs respectively. Moreover the fiber-induced dual graphs also represent the conne
ctivity information between vessels caused by fiber deposition. A primal-dual gr
aph convolution module is designed to facilitate the learning of spatial relatio
nships between vessels and fibers allowing for the joint exploration and interac
tion of their micro-structures. Experiments conducted on two datasets have shown
 that explicitly modeling the relationship between vessels and fibers leads to i
mproved fibrosis staging and enhanced interpretability.
****
UnO: Unsupervised Occupancy Fields for Perception and Forecasting
Ben Agro, Quinlan Sykora, Sergio Casas, Thomas Gilles, Raquel Urtasun; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2024, pp. 14487-14496
Perceiving the world and forecasting its future state is a critical task for sel
f-driving. Supervised approaches leverage annotated object labels to learn a mod
el of the world --- traditionally with object detections and trajectory predicti
ons or temporal bird's-eye-view (BEV) occupancy fields. However these annotation
s are expensive and typically limited to a set of predefined categories that do
not cover everything we might encounter on the road. Instead we learn to perceiv
e and forecast a continuous 4D (spatio-temporal) occupancy field with self-super
vision from LiDAR data. This unsupervised world model can be easily and effectiv
ely transferred to downstream tasks. We tackle point cloud forecasting by adding
 a lightweight learned renderer and achieve state-of-the-art performance in Argo
verse 2 nuScenes and KITTI. To further showcase its transferability we fine-tune
 our model for BEV semantic occupancy forecasting and show that it outperforms t
he fully supervised state-of-the-art especially when labeled data is scarce. Fin
ally when compared to prior state-of-the-art on spatio-temporal geometric occupa
ncy prediction our 4D world model achieves a much higher recall of objects from
classes relevant to self-driving.
****
SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities
Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas,
 Fei Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2024, pp. 14455-14465
Understanding and reasoning about spatial relationships is crucial for Visual Qu
estion Answering (VQA) and robotics. Vision Language Models (VLMs) have shown im
pressive performance in some VQA benchmarks but struggle with 3D spatial reasoni
ng such as recognizing distances or size differences between physical objects. T
his limitation may stem from a lack of 3D spatial knowledge in their training da
ta. To address this we propose training VLMs with extensive spatial reasoning da
ta from the internet. Our approach includes developing an automatic 3D spatial V
QA data generation framework capable of creating 2 billion VQA examples from 10
million real-world images. We explore various factors in the training process su
ch as data quality training pipeline and VLM architecture. Our work introduces t
he first Internet-scale 3D spatial reasoning dataset in metric space. By co-trai

ning a VLM with this dataset we significantly improve its performance in both qu alitative and quantitative spatial VQA. Additionally this enhanced VLM enables n ew applications in chain-of-thought spatial reasoning and robotics particularly in quantitative estimation.
****
InstructDiffusion: A Generalist Modeling Interface for Vision Tasks
Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, Dong Chen, Baining Guo; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12709-12720
We present InstructDiffusion a unified and generic framework for aligning comput er vision tasks with human instructions. Unlike existing approaches that integra te prior knowledge and pre-define the output space (e.g. categories and coordina tes) for each vision task we cast diverse vision tasks into a human-intuitive im age-manipulating process whose output space is a flexible and interactive pixel space. Concretely the model is built upon the diffusion process and is trained t o predict pixels according to user instructions such as encircling the man's lef t shoulder in red or applying a blue mask to the left car. InstructDiffusion cou ld handle a variety of vision tasks including understanding tasks (such as segme ntation and keypoint detection) and generative tasks (such as editing and enhanc ement) and outperforms prior methods on novel datasets. This represents a solid step towards a generalist modeling interface for vision tasks advancing artifici al general intelligence in the field of computer vision.
****
DreamVideo: Composing Your Dream Videos with Customized Subject and Motion
Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Z hang, Jingren Zhou, Hongming Shan; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2024, pp. 6537-6549
Customized generation using diffusion models has made impressive progress in ima ge generation but remains unsatisfactory in the challenging video generation tas k as it requires the controllability of both subjects and motions. To that end w e present DreamVideo a novel approach to generating personalized videos from a f ew static images of the desired subject and a few videos of target motion. Dream Video decouples this task into two stages subject learning and motion learning b y leveraging a pre-trained video diffusion model. The subject learning aims to a ccurately capture the fine appearance of the subject from provided images which is achieved by combining textual inversion and fine-tuning of our carefully desi gned identity adapter. In motion learning we architect a motion adapter and fine -tune it on the given videos to effectively model the target motion pattern. Com bining these two lightweight and efficient adapters allows for flexible customiz ation of any subject with any motion. Extensive experimental results demonstrate the superior performance of our DreamVideo over the state-of-the-art methods fo r customized video generation. Our project page is at https://dreamvideo-t2v.git hub.io.
****
Gated Fields: Learning Scene Reconstruction from Gated Videos
Andrea Ramazzina, Stefanie Walz, Pragyan Dahal, Mario Bijelic, Felix Heide; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2024, pp. 10530-10541
Reconstructing outdoor 3D scenes from temporal observations is a challenge that recent work on neural fields has offered a new avenue for. However existing meth ods that recover scene properties such as geometry appearance or radiance solely from RGB captures often fail when handling poorly-lit or texture-deficient regi ons. Similarly recovering scenes with scanning lidar sensors is also difficult d ue to their low angular sampling rate which makes recovering expansive real-worl d scenes difficult. Tackling these gaps we introduce Gated Fields - a neural sce ne reconstruction method that utilizes active gated video sequences. To this end we propose a neural rendering approach that seamlessly incorporates time-gated capture and illumination. Our method exploits the intrinsic depth cues in the ga ted videos achieving precise and dense geometry reconstruction irrespective of a

mbient illumination conditions. We validate the method across day and night scenarios and find that Gated Fields compares favorably to RGB and LiDAR reconstruction methods
****
RadarDistill: Boosting Radar-based Object Detection Performance via Knowledge Distillation from LiDAR Features
Geonho Bang, Kwangjin Choi, Jisong Kim, Dongsuk Kum, Jun Won Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15491-15500
The inherent noisy and sparse characteristics of radar data pose challenges in finding effective representations for 3D object detection. In this paper we propose RadarDistill a novel knowledge distillation (KD) method which can improve the representation of radar data by leveraging LiDAR data. RadarDistill successfully transfers desirable characteristics of LiDAR features into radar features using three key components: Cross-Modality Alignment (CMA) Activation-based Feature Distillation (AFD) and Proposal-based Feature Distillation (PFD). CMA enhances the density of radar features by employing multiple layers of dilation operations effectively addressing the challenge of inefficient knowledge transfer from LiDAR to radar. AFD selectively transfers knowledge based on regions of the LiDAR features with a specific focus on areas where activation intensity exceeds a predefined threshold. PFD similarly guides the radar network to selectively mimic features from the LiDAR network within the object proposals. Our comparative analyses conducted on the nuScenes datasets demonstrate that RadarDistill achieves state-of-the-art (SOTA) performance for radar-only object detection task recording 20.5% in mAP and 43.7% in NDS. Also RadarDistill significantly improves the performance of the camera-radar fusion model.
****
Probabilistic Sampling of Balanced K-Means using Adiabatic Quantum Computing
Jan-Nico Zaech, Martin Danelljan, Tolga Birdal, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26191-26201
Adiabatic quantum computing (AQC) is a promising approach for discrete and often NP-hard optimization problems. Current AQCs allow to implement problems of research interest which has sparked the development of quantum representations for many computer vision tasks. Despite requiring multiple measurements from the noisy AQC current approaches only utilize the best measurement discarding information contained in the remaining ones. In this work we explore the potential of using this information for probabilistic balanced k-means clustering. Instead of discarding non-optimal solutions we propose to use them to compute calibrated posterior probabilities with little additional compute cost. This allows us to identify ambiguous solutions and data points which we demonstrate on a D-Wave AQC on synthetic tasks and real visual data.
****
UniPT: Universal Parallel Tuning for Transfer Learning with Efficient Parameter and Memory
Haiwen Diao, Bo Wan, Ying Zhang, Xu Jia, Huchuan Lu, Long Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28729-28740
Parameter-efficient transfer learning (PETL) i.e. fine-tuning a small portion of parameters is an effective strategy for adapting pre-trained models to downstream domains. To further reduce the memory demand recent PETL works focus on the more valuable memory-efficient characteristic. In this paper we argue that the scalability adaptability and generalizability of state-of-the-art methods are hindered by structural dependency and pertinency on specific pre-trained backbones. To this end we propose a new memory-efficient PETL strategy Universal Parallel Tuning (UniPT) to mitigate these weaknesses. Specifically we facilitate the transfer process via a lightweight and learnable parallel network which consists of: 1) A parallel interaction module that decouples the sequential connections and processes the intermediate activations detachedly from the pre-trained network. 2) A confidence aggregation module that learns optimal strategies adaptively for

integrating cross-layer features. We evaluate UniPT with different backbones (e.g. T5 VSEinfinity CLIP4Clip Clip-ViL and MDETR) on various vision-and-language and pure NLP tasks. Extensive ablations on 18 datasets have validated that UniPT can not only dramatically reduce memory consumption and outperform the best competitor but also achieve competitive performance over other plain PETL methods with lower training memory overhead. Our code is publicly available at: https://github.com/Paranioar/UniPT.
****
Composed Video Retrieval via Enriched Context and Discriminative Embeddings
Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26896-26906
Composed video retrieval (CoVR) is a challenging prob- lem in computer vision which has recently highlighted the in- tegration of modification text with visual queries for more so- phisticated video search in large databases. Existing works predominantly rely on visual queries combined with modi- fication text to distinguish relevant videos. However such a strategy struggles to fully preserve the rich query-specific context in retrieved target videos and only represents the target video using visual embedding. We introduce a novel CoVR framework that leverages detailed language descrip- tions to explicitly encode query-specific contextual informa- tion and learns discriminative embeddings of vision only text only and vision-text for better alignment to accurately retrieve matched target videos. Our proposed framework can be flexibly employed for both composed video (CoVR) and image (CoIR) retrieval tasks. Experiments on three datasets show that our approach obtains state-of-the-art per- formance for both CovR and zero-shot CoIR tasks achiev- ing gains as high as around 7% in terms of recall@K=1 score. Our code detailed language descriptions for WebViD- CoVR dataset are available at https://github.com/ OmkarThawakar/composed-video-retrieval.
****
Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model
Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, Xiu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8941-8951
Using reinforcement learning with human feedback (RLHF) has shown significant promise in fine-tuning diffusion models. Previous methods start by training a reward model that aligns with human preferences then leverage RL techniques to fine-tune the underlying models. However crafting an efficient reward model demands extensive datasets optimal architecture and manual hyperparameter tuning making the process both time and cost-intensive. The direct preference optimization (DPO) method effective in fine-tuning large language models eliminates the necessity for a reward model. However the extensive GPU memory requirement of the diffusion model's denoising process hinders the direct application of the DPO method. To address this issue we introduce the Direct Preference for Denoising Diffusion Policy Optimization (D3PO) method to directly fine-tune diffusion models. The theoretical analysis demonstrates that although D3PO omits training a reward model it effectively functions as the optimal reward model trained using human feedback data to guide the learning process. This approach requires no training of a reward model proving to be more direct cost-effective and minimizing computational overhead. In experiments our method uses the relative scale of objectives as a proxy for human preference delivering comparable results to methods using ground-truth rewards. Moreover D3PO demonstrates the ability to reduce image distortion rates and generate safer images overcoming challenges lacking robust reward models. Our code is publicly available at https://github.com/yk7333/D3PO.
****
Perceptual Assessment and Optimization of HDR Image Rendering
Peibei Cao, Rafal K. Mantiuk, Kede Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22433-22443
High dynamic range (HDR) rendering has the ability to faithfully reproduce the wide luminance ranges in natural scenes but how to accurately assess the rendering quality is relatively underexplored. Existing quality models are mostly design

ed for low dynamic range (LDR) images and do not align well with human perception of HDR image quality. To fill this gap we propose a family of HDR quality metrics in which the key step is employing a simple inverse display model to decompose an HDR image into a stack of LDR images with varying exposures. Subsequently these decomposed images are assessed through well-established LDR quality metrics. Our HDR quality models present three distinct benefits. First they directly inherit the recent advancements of LDR quality metrics. Second they do not rely on human perceptual data of HDR image quality for re-calibration. Third they facilitate the alignment and prioritization of specific luminance ranges for more accurate and detailed quality assessment. Experimental results show that our HDR quality metrics consistently outperform existing models in terms of quality assessment on four HDR image quality datasets and perceptual optimization of HDR novel view synthesis.
****
Multiview Aerial Visual RECognition (MAVREC): Can Multi-view Improve Aerial Visual Perception?
Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatsubhra Chakraborty, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22678-22690
Despite the commercial abundance of UAVs aerial data acquisition remains challenging and the existing Asia and North America-centric open-source UAV datasets are small-scale or low-resolution and lack diversity in scene contextuality. Additionally the color content of the scenes solar zenith angle and population density of different geographies influence the data diversity. These factors conjointly render suboptimal aerial-visual perception of the deep neural network (DNN) models trained primarily on the ground view data including the open-world foundational models. To pave the way for a transformative era of aerial detection we present Multiview Aerial Visual RECognition (MAVREC) a video dataset where we record synchronized scenes from different perspectives --- ground camera and drone-mounted camera. MAVREC consists of around 2.5 hours of industry-standard 2.7K resolution video sequences more than 0.5 million frames and 1.1 million annotated bounding boxes. This makes MAVREC the largest ground and aerial view dataset and the fourth largest among all drone-based datasets across all modalities and tasks. Through our extensive benchmarking on MAVREC we recognize that augmenting object detectors with ground view images from the corresponding geographical location is a superior pre-training strategy for aerial detection. Building on this strategy we benchmark MAVREC with a curriculum-based semi-supervised object detection approach that leverages labeled (ground and aerial) and unlabeled (only aerial) images to enhance aerial detection.
****
Diffusion-driven GAN Inversion for Multi-Modal Face Image Generation
Jihyun Kim, Changjae Oh, Hoseok Do, Soohyun Kim, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10403-10412
We present a new multi-modal face image generation method that converts a text prompt and a visual input such as a semantic mask or scribble map into a photo-realistic face image. To do this we combine the strengths of Generative Adversarial networks (GANs) and diffusion models (DMs) by employing the multi-modal features in the DM into the latent space of the pre-trained GANs. We present a simple mapping and a style modulation network to link two models and convert meaningful representations in feature maps and attention maps into latent codes. With GAN inversion the estimated latent codes can be used to generate 2D or 3D-aware facial images. We further present a multi-step training strategy that reflects textual and structural representations into the generated image. Our proposed network produces realistic 2D multi-view and stylized face images which align well with inputs. We validate our method by using pre-trained 2D and 3D GANs and our results outperform existing methods. Our project page is available at https://github.com/1211sh/Diffusiondriven_GAN-Inversion/.
****
Low-Rank Knowledge Decomposition for Medical Foundation Models

Yuhang Zhou, Haolin Li, Siyuan Du, Jiangchao Yao, Ya Zhang, Yanfeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11611-11620
The popularity of large-scale pre-training has promoted the development of medical foundation models. However some studies have shown that although foundation models exhibit strong general feature extraction capabilities their performance on specific tasks is still inferior to task-specific methods. In this paper we explore a new perspective called "Knowledge Decomposition" to improve the performance on specific medical tasks which deconstruct the foundation model into multiple lightweight expert models each dedicated to a particular task with the goal of improving specialization while concurrently mitigating resource expenditure. To accomplish the above objective we design a novel framework named Low-Rank Knowledge Decomposition (LoRKD) which explicitly separates graidents by incorporating low-rank expert modules and the efficient knowledge separation convolution. Extensive experimental results demonstrate that the decomposed models perform well in terms of performance and transferability even surpassing the original foundation models. Source code is available at: https://github.com/MediaBrain-SJTU/LoRKD
****
SaCo Loss: Sample-wise Affinity Consistency for Vision-Language Pre-training
Sitong Wu, Haoru Tan, Zhuotao Tian, Yukang Chen, Xiaojuan Qi, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27358-27369
Vision-language pre-training (VLP) aims to learn joint representations of vision and language modalities. The contrastive paradigm is currently dominant in this field. However we observe a notable misalignment phenomenon that is the affinity between samples has an obvious disparity across different modalities namely "Affinity Inconsistency Problem". Our intuition is that for a well-aligned model two images that look similar to each other should have the same level of similarity as their corresponding texts that describe them. In this paper we first investigate the reason of this inconsistency problem. We discover that the lack of consideration for sample-wise affinity consistency across modalities in existing training objectives is the central cause. To address this problem we propose a novel loss function named Sample-wise affinity Consistency (SaCo) loss which is designed to enhance such consistency by minimizing the distance between image embedding similarity and text embedding similarity for any two samples. Our SaCo loss can be easily incorporated into existing vision-language models as an additional loss due to its complementarity for most training objectives. In addition considering that pre-training from scratch is computationally expensive we also provide a more efficient way to continuously pre-train on a converged model by integrating our loss. Experimentally the model trained with our SaCo loss significantly outperforms the baseline on a variety of vision and language tasks.
****
Steganographic Passport: An Owner and User Verifiable Credential for Deep Model IP Protection Without Retraining
Qi Cui, Ruohan Meng, Chaohui Xu, Chip-Hong Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12302-12311
Ensuring the legal usage of deep models is crucial to promoting trustable accountable and responsible artificial intelligence innovation. Current passport-based methods that obfuscate model functionality for license-to-use and ownership verifications suffer from capacity and quality constraints as they require retraining the owner model for new users. They are also vulnerable to advanced Expanded Residual Block ambiguity attacks. We propose Steganographic Passport which uses an invertible steganographic network to decouple license-to-use from ownership verification by hiding the user's identity images into the owner-side passport and recovering them from their respective user-side passports. An irreversible and collision-resistant hash function is used to avoid exposing the owner-side passport from the derived user-side passports and increase the uniqueness of the model signature. To safeguard both the passport and model's weights against advance

d ambiguity attacks an activation-level obfuscation is proposed for the verifica
tion branch of the owner's model. By jointly training the verification and deplo
yment branches their weights become tightly coupled. The proposed method support
s agile licensing of deep models by providing a strong ownership proof and licen
se accountability without requiring a separate model retraining for the admissio
n of every new user. Experiment results show that our Steganographic Passport ou
tperforms other passport-based deep model protection methods in robustness again
st various known attacks.
****
Stable Neighbor Denoising for Source-free Domain Adaptive Segmentation
Dong Zhao, Shuang Wang, Qi Zang, Licheng Jiao, Nicu Sebe, Zhun Zhong; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 23416-23427
We study source-free unsupervised domain adaptation (SFUDA) for semantic segment
ation which aims to adapt a source-trained model to the target domain without ac
cessing the source data. Many works have been proposed to address this challengi
ng problem among which uncertainty based self-training is a predominant approach
. However without comprehensive denoising mechanisms they still largely fall int
o biased estimates when dealing with different domains and confirmation bias. In
 this paper we observe that pseudo-label noise is mainly contained in unstable s
amples in which the predictions of most pixels undergo significant variations du
ring self-training. Inspired by this we propose a novel mechanism to denoise uns
table samples with stable ones. Specifically we introduce the Stable Neighbor De
noising (SND) approach which effectively discovers highly correlated stable and
unstable samples by nearest neighbor retrieval and guides the reliable optimizat
ion of unstable samples by bi-level learning. Moreover we compensate for the sta
ble set by object-level object paste which can further eliminate the bias caused
 by less learned classes. Our SND enjoys two advantages. First SND does not requ
ire a specific segmentor structure endowing its universality. Second SND simulta
neously addresses the issues of class domain and confirmation biases during adap
tation ensuring its effectiveness. Extensive experiments show that SND consisten
tly outperforms state-of-the-art methods in various SFUDA semantic segmentation
settings. In addition SND can be easily integrated with other approaches obtaini
ng further improvements. The source code will be publicly available.
****
SynSP: Synergy of Smoothness and Precision in Pose Sequences Refinement
Tao Wang, Lei Jin, Zheng Wang, Jianshu Li, Liang Li, Fang Zhao, Yu Cheng, Li Yua
n, Li Zhou, Junliang Xing, Jian Zhao; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1824-1833
Predicting human pose sequences via existing pose estimators often encounters va
rious estimation errors. Motion refinement methods aim to optimize the predicted
 human pose sequences from pose estimators while ensuring minimal computational
overhead and latency. Prior investigations have primarily concentrated on striki
ng a balance between the two objectives i.e. smoothness and precision while opti
mizing the predicted pose sequences. However it has come to our attention that t
he tension between these two objectives can provide additional quality cues abou
t the predicted pose sequences. These cues in turn are able to aid the network i
n optimizing lower-quality poses. To leverage this quality information we propos
e a motion refinement network termed SynSP to achieve a Synergy of Smoothness an
d Precision in the sequence refinement tasks. Moreover SynSP can also address mu
lti-view poses of one person simultaneously fixing inaccuracies in predicted pos
es through heightened attention to similar poses from other views thereby amplif
ying the resultant quality cues and overall performance. Compared with previous
methods SynSP benefits from both pose quality and multi-view information with a
much shorter input sequence length achieving state-of-the-art results among four
 challenging datasets involving 2D 3D and SMPL pose representations in both sing
le-view and multi-view scenes. Github code: https://github.com/InvertedForest/Sy
nSP.
****
En3D: An Enhanced Generative Model for Sculpting 3D Humans from 2D Synthetic Dat

a
Yifang Men, Biwen Lei, Yuan Yao, Miaomiao Cui, Zhouhui Lian, Xuansong Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9981-9991
We present En3D an enhanced generative scheme for sculpting high-quality 3D human avatars. Unlike previous works that rely on scarce 3D datasets or limited 2D collections with imbalanced viewing angles and imprecise pose priors our approach aims to develop a zero-shot 3D generative scheme capable of producing visually realistic geometrically accurate and content-wise diverse 3D humans without relying on pre-existing 3D or 2D assets. To address this challenge we introduce a meticulously crafted workflow that implements accurate physical modeling to learn the enhanced 3D generative model from synthetic 2D data. During inference we integrate optimization modules to bridge the gap between realistic appearances and coarse 3D shapes. Specifically En3D comprises three modules: a 3D generator that accurately models generalizable 3D humans with realistic appearance from synthesized balanced diverse and structured human images; a geometry sculptor that enhances shape quality using multi-view normal constraints for intricate human structure; and a texturing module that disentangles explicit texture maps with fidelity and editability leveraging semantical UV partitioning and a differentiable rasterizer. Experimental results show that our approach significantly outperforms prior works in terms of image quality geometry accuracy and content diversity. We also showcase the applicability of our generated avatars for animation and editing as well as the scalability of our approach for content-style free adaptation.
****
Neural Visibility Field for Uncertainty-Driven Active Mapping
Shangjie Xue, Jesse Dill, Pranay Mathur, Frank Dellaert, Panagiotis Tsiotra, Danfei Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18122-18132
This paper presents Neural Visibility Field (NVF) a novel uncertainty quantification method for Neural Radiance Fields (NeRF) applied to active mapping. Our key insight is that regions not visible in the training views lead to inherently unreliable color predictions by NeRF at this region resulting in increased uncertainty in the synthesized views. To address this we propose to use Bayesian Networks to composite position-based field uncertainty into ray-based uncertainty in camera observations. Consequently NVF naturally assigns higher uncertainty to unobserved regions aiding robots to select the most informative next viewpoints. Extensive evaluations show that NVF excels not only in uncertainty quantification but also in scene reconstruction for active mapping outperforming existing methods. More details can be found at https://sites.google.com/view/nvf-cvpr24/.
****
Tri-Perspective View Decomposition for Geometry-Aware Depth Completion
Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4874-4884
Depth completion is a vital task for autonomous driving as it involves reconstructing the precise 3D geometry of a scene from sparse and noisy depth measurements. However most existing methods either rely only on 2D depth representations or directly incorporate raw 3D point clouds for compensation which are still insufficient to capture the fine-grained 3D geometry of the scene. To address this challenge we introduce Tri-Perspective View Decomposition (TPVD) a novel framework that can explicitly model 3D geometry. In particular (1) TPVD ingeniously decomposes the original point cloud into three 2D views one of which corresponds to the sparse depth input. (2) We design TPV Fusion to update the 2D TPV features through recurrent 2D-3D-2D aggregation where a Distance-Aware Spherical Convolution (DASC) is applied. (3) By adaptively choosing TPV affinitive neighbors the newly proposed Geometric Spatial Propagation Network (GSPN) further improves the geometric consistency. As a result our TPVD outperforms existing methods on KITTI NYUv2 and SUN RGBD. Furthermore we build a novel depth completion dataset named TOFDC which is acquired by the time-of-flight (TOF) sensor and the color camera

on smartphones.
****
Boosting Adversarial Training via Fisher-Rao Norm-based Regularization
Xiangyu Yin, Wenjie Ruan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24544-24553
Adversarial training is extensively utilized to improve the adversarial robustness of deep neural networks. Yet mitigating the degradation of standard generalization performance in adversarial-trained models remains an open problem. This paper attempts to resolve this issue through the lens of model complexity. First We leverage the Fisher-Rao norm a geometrically invariant metric for model complexity to establish the non-trivial bounds of the Cross-Entropy Loss-based Rademacher complexity for a ReLU-activated Multi-Layer Perceptron. Building upon this observation we propose a novel regularization framework called Logit-Oriented Adversarial Training (LOAT) which can mitigate the trade-off between robustness and accuracy while imposing only a negligible increase in computational overhead. Our extensive experiments demonstrate that the proposed regularization strategy can boost the performance of the prevalent adversarial training algorithms including PGD-AT TRADES TRADES (LSE) MART and DM-AT across various network architectures. Our code will be available at https://github.com/TrustAI/LOAT.
****
Learned Representation-Guided Diffusion Models for Large-Image Generation
Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, Dimitris Samaras; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8532-8542
To synthesize high-fidelity samples diffusion models typically require auxiliary data to guide the generation process. However it is impractical to procure the painstaking patch-level annotation effort required in specialized domains like histopathology and satellite imagery; it is often performed by domain experts and involves hundreds of millions of patches. Modern-day self-supervised learning (SSL) representations encode rich semantic and visual information. In this paper we posit that such representations are expressive enough to act as proxies to fine-grained human labels. We introduce a novel approach that trains diffusion models conditioned on embeddings from SSL. Our diffusion models successfully project these features back to high-quality histopathology and remote sensing images. In addition we construct larger images by assembling spatially consistent patches inferred from SSL embeddings preserving long-range dependencies. Augmenting real data by generating variations of real images improves downstream classifier accuracy for patch-level and larger image-scale classification tasks. Our models are effective even on datasets not encountered during training demonstrating their robustness and generalizability. Generating images from learned embeddings is agnostic to the source of the embeddings. The SSL embeddings used to generate a large image can either be extracted from a reference image or sampled from an auxiliary model conditioned on any related modality (e.g. class labels text genomic data). As proof of concept we introduce the text-to-large image synthesis paradigm where we successfully synthesize large pathology and satellite images out of text descriptions.
****
DAVE - A Detect-and-Verify Paradigm for Low-Shot Counting
Jer Pelhan, Alan Lukeži?, Vitjan Zavrtanik, Matej Kristan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23293-23302
Low-shot counters estimate the number of objects corresponding to a selected category based on only few or no exemplars annotated in the image. The current state-of-the-art estimates the total counts as the sum over the object location density map but do not provide object locations and sizes which are crucial for many applications. This is addressed by detection-based counters which however fall behind in the total count accuracy. Furthermore both approaches tend to overestimate the counts in the presence of other object classes due to many false positives. We propose DAVE a low-shot counter based on a detect-and-verify paradigm that avoids the aforementioned issues by first generating a high-recall detection

set and then verifying the detections to identify and remove the outliers.This j ointly increases the recall and precision leading to accurate counts. DAVE outpe rforms the top density-based counters by ?20% in the total count MAE it outperfo rms the most recent detection-based counter by ?20% in detection quality and set s a new state-of-the-art in zero-shot as well as text-prompt-based counting.
****
Ranni: Taming Text-to-Image Diffusion for Accurate Instruction Following
Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, Jingren Zhou; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 4744-4753
Existing text-to-image (T2I) diffusion models usually struggle in interpreting c omplex prompts especially those with quantity object-attribute binding and multi -subject descriptions. In this work we introduce a semantic panel as the middlew are in decoding texts to images supporting the generator to better follow instru ctions. The panel is obtained through arranging the visual concepts parsed from the input text by the aid of large language models and then injected into the de noising network as a detailed control signal to complement the text condition. T o facilitate text-to-panel learning we come up with a carefully designed semanti c formatting protocol accompanied by a fully-automatic data preparation pipeline . Thanks to such a design our approach which we call Ranni manages to enhance a pre-trained T2I generator regarding its textual controllability. More importantl y the introduction of the generative middleware brings a more convenient form of interaction (i.e. directly adjusting the elements in the panel or using languag e instructions) and further allows users to finely customize their generation ba sed on which we develop a practical system and showcase its potential in continu ous generation and chatting-based editing.
****
Relaxed Contrastive Learning for Federated Learning
Seonguk Seo, Jinkyu Kim, Geeho Kim, Bohyung Han; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12279-12288
We propose a novel contrastive learning framework to effectively address the cha llenges of data heterogeneity in federated learning. We first analyze the incons istency of gradient updates across clients during local training and establish i ts dependence on the distribution of feature representations leading to the deri vation of the supervised contrastive learning (SCL) objective to mitigate local deviations. In addition we show that a naive integration of SCL into federated l earning incurs representation collapse resulting in slow convergence and limited performance gains. To address this issue we introduce a relaxed contrastive lea rning loss that imposes a divergence penalty on excessively similar sample pairs within each class. This strategy prevents collapsed representations and enhance s feature transferability facilitating collaborative training and leading to sig nificant performance improvements. Our framework outperforms all existing federa ted learning approaches by significant margins on the standard benchmarks as dem onstrated by extensive experimental results. The source code is available at our project page(https://github.com/skynbe/FedRCL).
****
Direct2.5: Diverse Text-to-3D Generation via Multi-view 2.5D Diffusion
Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8744-8753
Recent advances in generative AI have unveiled significant potential for the cre ation of 3D content. However current methods either apply a pre-trained 2D diffu sion model with the time-consuming score distillation sampling (SDS) or a direct 3D diffusion model trained on limited 3D data losing generation diversity. In t his work we approach the problem by employing a multi-view 2.5D diffusion fine-t uned from a pre-trained 2D diffusion model. The multi-view 2.5D diffusion direct ly models the structural distribution of 3D data while still maintaining the str ong generalization ability of the original 2D diffusion model filling the gap be tween 2D diffusion-based and direct 3D diffusion-based methods for 3D content ge neration. During inference multi-view normal maps are generated using the 2.5D d

iffusion and a novel differentiable rasterization scheme is introduced to fuse the almost consistent multi-view normal maps into a consistent 3D model. We further design a normal-conditioned multi-view image generation module for fast appearance generation given the 3D geometry. Our method is a one-pass diffusion process and does not require any SDS optimization as post-processing. We demonstrate through extensive experiments that our direct 2.5D generation with the specially-designed fusion scheme can achieve diverse mode-seeking-free and high-fidelity 3D content generation in only 10 seconds.
****
Efficient LoFTR: Semi-Dense Local Feature Matching with Sparse-Like Speed
Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21666-21675
We present a novel method for efficiently producing semi-dense matches across images. Previous detector-free matcher LoFTR has shown remarkable matching capability in handling large-viewpoint change and texture-poor scenarios but suffers from low efficiency. We revisit its design choices and derive multiple improvements for both efficiency and accuracy. One key observation is that performing the transformer over the entire feature map is redundant due to shared local information therefore we propose an aggregated attention mechanism with adaptive token selection for efficiency. Furthermore we find spatial variance exists in LoFTR's fine correlation module which is adverse to matching accuracy. A novel two-stage correlation layer is proposed to achieve accurate subpixel correspondences for accuracy improvement. Our efficiency optimized model is ~ 2.5xfaster than LoFTR which can even surpass state-of-the-art efficient sparse matching pipeline SuperPoint + LightGlue. Moreover extensive experiments show that our method can achieve higher accuracy compared with competitive semi-dense matchers with considerable efficiency benefits. This opens up exciting prospects for large-scale or latency-sensitive applications such as image retrieval and 3D reconstruction. Project page: https://zju3dv.github.io/efficientloftr/.
****
Contextual Augmented Global Contrast for Multimodal Intent Recognition
Kaili Sun, Zhiwen Xie, Mang Ye, Huyin Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26963-26973
Multimodal intent recognition (MIR) aims to perceive the human intent polarity via language visual and acoustic modalities. The inherent intent ambiguity makes it challenging to recognize in multimodal scenarios. Existing MIR methods tend to model the individual video independently ignoring global contextual information across videos. This learning manner inevitably introduces perception biases exacerbated by the inconsistencies of the multimodal representation amplifying the intent uncertainty. This challenge motivates us to explore effective global context modeling. Thus we propose a context-augmented global contrast (CAGC) method to capture rich global context features by mining both intra-and cross-video context interactions for MIR. Concretely we design a context-augmented transformer module to extract global context dependencies across videos. To further alleviate error accumulation and interference we develop a cross-video bank that retrieves effective video sources by considering both intentional tendency and video similarity. Furthermore we introduce a global context-guided contrastive learning scheme designed to mitigate inconsistencies arising from global context and individual modalities in different feature spaces. This scheme incorporates global cues as the supervision to capture robust the multimodal intent representation. Experiments demonstrate CAGC obtains superior performance than state-of-the-art MIR methods. We also generalize our approach to a closely related task multimodal sentiment analysis achieving the comparable performance.
****
Pre-trained Model Guided Fine-Tuning for Zero-Shot Adversarial Robustness
Sibo Wang, Jie Zhang, Zheng Yuan, Shiguang Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24502-24511
Large-scale pre-trained vision-language models like CLIP have demonstrated impressive performance across various tasks and exhibit remarkable zero-shot generali

zation capability while they are also vulnerable to imperceptible adversarial ex
amples. Existing works typically employ adversarial training (fine-tuning) as a
defense method against adversarial examples. However direct application to the C
LIP model may result in overfitting compromising the model's capacity for genera
lization. In this paper we propose Pre-trained Model Guided Adversarial Fine-Tun
ing (PMG-AFT) method which leverages supervision from the original pre-trained m
odel by carefully designing an auxiliary branch to enhance the model's zero-shot
 adversarial robustness. Specifically PMG-AFT minimizes the distance between the
 features of adversarial examples in the target model and those in the pre-train
ed model aiming to preserve the generalization features already captured by the
pre-trained model. Extensive Experiments on 15 zero-shot datasets demonstrate th
at PMG-AFT significantly outperforms the state-of-the-art method improving the t
op-1 robust accuracy by an average of 4.99%. Furthermore our approach consistent
ly improves clean accuracy by an average of 8.72%.
****
MatFuse: Controllable Material Generation with Diffusion Models
Giuseppe Vecchio, Renato Sortino, Simone Palazzo, Concetto Spampinato; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2024, pp. 4429-4438
Creating high-quality materials in computer graphics is a challenging and time-c
onsuming task which requires great expertise. To simplify this process we introd
uce MatFuse a unified approach that harnesses the generative power of diffusion
models for creation and editing of 3D materials. Our method integrates multiple
sources of conditioning including color palettes sketches text and pictures enha
ncing creative possibilities and granting fine-grained control over material syn
thesis. Additionally MatFuse enables map-level material editing capabilities thr
ough latent manipulation by means of a multi-encoder compression model which lea
rns a disentangled latent representation for each map. We demonstrate the effect
iveness of MatFuse under multiple conditioning settings and explore the potentia
l of material editing. Finally we assess the quality of the generated materials
both quantitatively in terms of CLIP-IQA and FID scores and qualitatively by con
ducting a user study. Source code for training MatFuse and supplemental material
s are publicly available at https://gvecchio.com/matfuse.
****
CoGS: Controllable Gaussian Splatting
Heng Yu, Joel Julin, Zoltán A. Milacski, Koichiro Niinuma, László A. Jeni; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 21624-21633
Capturing and re-animating the 3D structure of articulated objects present signi
ficant barriers. On one hand methods requiring extensively calibrated multi-view
 setups are prohibitively complex and resource-intensive limiting their practica
l applicability. On the other hand while single-camera Neural Radiance Fields (N
eRFs) offer a more streamlined approach they have excessive training and renderi
ng costs. 3D Gaussian Splatting would be a suitable alternative but for two reas
ons. Firstly existing methods for 3D dynamic Gaussians require synchronized mult
i-view cameras and secondly the lack of controllability in dynamic scenarios. We
 present CoGS a method for Controllable Gaussian Splatting that enables the dire
ct manipulation of scene elements offering real-time control of dynamic scenes w
ithout the prerequisite of pre-computing control signals. We evaluated CoGS usin
g both synthetic and real-world datasets that include dynamic objects that diffe
r in degree of difficulty. In our evaluations CoGS consistently outperformed exi
sting dynamic and controllable neural representations in terms of visual fidelit
y.
****
Partial-to-Partial Shape Matching with Geometric Consistency
Viktoria Ehm, Maolin Gao, Paul Roetzer, Marvin Eisenberger, Daniel Cremers, Flor
ian Bernard; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2024, pp. 27488-27497
Finding correspondences between 3D shapes is an important and long-standing prob
lem in computer vision graphics and beyond. A prominent challenge are partial-to

-partial shape matching settings which occur when the shapes to match are only o
bserved incompletely (e.g. from 3D scanning). Although partial-to-partial matchi
ng is a highly relevant setting in practice it is rarely explored. Our work brid
ges the gap between existing (rather artificial) 3D full shape matching and part
ial-to-partial real-world settings by exploiting geometric consistency as a stro
ng constraint. We demonstrate that it is indeed possible to solve this challengi
ng problem in a variety of settings. For the first time we achieve geometric con
sistency for partial-to-partial matching which is realized by a novel integer no
n-linear program formalism building on triangle product spaces along with a new
pruning algorithm based on linear integer programming. Further we generate a new
 inter-class dataset for partial-to-partial shape-matching. We show that our met
hod outperforms current SOTA methods on both an established intra-class dataset
and our novel inter-class dataset.
****
Descriptor and Word Soups: Overcoming the Parameter Efficiency Accuracy Tradeoff
 for Out-of-Distribution Few-shot Learning
Christopher Liao, Theodoros Tsiligkaridis, Brian Kulis; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2701
5-27025
Over the past year a large body of multimodal research has emerged around zero-s
hot evaluation using GPT descriptors. These studies boost the zero-shot accuracy
 of pretrained VL models with an ensemble of label-specific text generated by GP
T. A recent study WaffleCLIP demonstrated that similar zero-shot accuracy can be
 achieved with an ensemble of random descriptors. However both zero-shot methods
 are un-trainable and consequently sub-optimal when some few-shot out-of-distrib
ution (OOD) training data is available. Inspired by these prior works we present
 two more flexible methods called descriptor and word soups which do not require
 an LLM at test time and can leverage training data to increase OOD target accur
acy. Descriptor soup greedily selects a small set of textual descriptors using g
eneric few-shot training data then calculates robust class embeddings using the
selected descriptors. Word soup greedily assembles a chain of words in a similar
 manner. Compared to existing few-shot soft prompt tuning methods word soup requ
ires fewer parameters by construction and less GPU memory since it does not requ
ire backpropagation. Both soups outperform current published few-shot methods ev
en when combined with SoTA zero-shot methods on cross-dataset and domain general
ization benchmarks. Compared with SoTA prompt and descriptor ensembling methods
such as ProDA and WaffleCLIP word soup achieves higher OOD accuracy with fewer e
nsemble members. Please checkout our code: https://github.com/Chris210634/word_s
oups
****
Harnessing the Power of MLLMs for Transferable Text-to-Image Person ReID
Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, Dapeng Tao; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 17127-17137
Text-to-image person re-identification (ReID) retrieves pedestrian images accord
ing to textual descriptions. Manually annotating textual descriptions is time-co
nsuming restricting the scale of existing datasets and therefore the generalizat
ion ability of ReID models. As a result we study the transferable text-to-image
ReID problem where we train a model on our proposed large-scale database and dir
ectly deploy it to various datasets for evaluation. We obtain substantial traini
ng data via Multi-modal Large Language Models (MLLMs). Moreover we identify and
address two key challenges in utilizing the obtained textual descriptions. First
 an MLLM tends to generate descriptions with similar structures causing the mode
l to overfit specific sentence patterns. Thus we propose a novel method that use
s MLLMs to caption images according to various templates. These templates are ob
tained using a multi-turn dialogue with a Large Language Model (LLM). Therefore
we can build a large-scale dataset with diverse textual descriptions. Second an
MLLM may produce incorrect descriptions. Hence we introduce a novel method that
automatically identifies words in a description that do not correspond with the
image. This method is based on the similarity between one text and all patch tok

en embeddings in the image. Then we mask these words with a larger probability i
n the subsequent training epoch alleviating the impact of noisy textual descript
ions. The experimental results demonstrate that our methods significantly boost
the direct transfer text-to-image ReID performance. Benefiting from the pre-trai
ned model weights we also achieve state-of-the-art performance in the traditiona
l evaluation settings.
****
360+x: A Panoptic Multi-modal Scene Understanding Dataset
Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, Jianbo Jiao; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 19373-19382
Human perception of the world is shaped by a multitude of viewpoints and modalit
ies. While many existing datasets focus on scene understanding from a certain pe
rspective (e.g. egocentric or third-person views) our dataset offers a panoptic
perspective (i.e. multiple viewpoints with multiple data modalities). Specifical
ly we encapsulate third-person panoramic and front views as well as egocentric m
onocular/binocular views with rich modalities including video multi-channel audi
o directional binaural delay location data and textual scene descriptions within
 each scene captured presenting comprehensive observation of the world. To the b
est of our knowledge this is the first database that covers multiple viewpoints
with multiple data modalities to mimic how daily information is accessed in the
real world. Through our benchmark analysis we presented 5 different scene unders
tanding tasks on the proposed 360+x dataset to evaluate the impact and benefit o
f each data modality and perspective in panoptic scene understanding. We hope th
is unique dataset could broaden the scope of comprehensive scene understanding a
nd encourage the community to approach these problems from more diverse perspect
ives.
****
Weakly Supervised Video Individual Counting
Xinyan Liu, Guorong Li, Yuankai Qi, Ziheng Yan, Zhenjun Han, Anton van den Henge
l, Ming-Hsuan Yang, Qingming Huang; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2024, pp. 19228-19237
Video Individual Counting (VIC) aims to predict the number of unique individuals
 in a single video. Existing methods learn representations based on trajectory l
abels for individuals which are annotation-expensive. To provide a more realisti
c reflection of the underlying practical challenge we introduce a weakly supervi
sed VIC task wherein trajectory labels are not provided. Instead two types of la
bels are provided to indicate traffic entering the field of view (inflow) and le
aving the field view (outflow). We also propose the first solution as a baseline
 that formulates the task as a weakly supervised contrastive learning problem un
der group-level matching. In doing so we devise an end-to-end trainable soft con
trastive loss to drive the network to distinguish inflow outflow and the remaini
ng. To facilitate future study in this direction we generate annotations from th
e existing VIC datasets SenseCrowd and CroHD and also build a new dataset UAVVIC
. Extensive results show that our baseline weakly supervised method outperforms
supervised methods and thus little information is lost in the transition to the
more practically relevant weakly supervised task. The code and trained model can
 be found at CGNet.
****
Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion
 Models
Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, Nenghai Yu; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 12162-12171
Ethical concerns surrounding copyright protection and inappropriate content gene
ration pose challenges for the practical implementation of diffusion models. One
 effective solution involves watermarking the generated images. However existing
 methods often compromise the model performance or require additional training w
hich is undesirable for operators and users. To address this issue we propose Ga
ussian Shading a diffusion model watermarking technique that is both performance

-lossless and training-free while serving the dual purpose of copyright protection and tracing of offending content. Our watermark embedding is free of model parameter modifications and thus is plug-and-play. We map the watermark to latent representations following a standard Gaussian distribution which is indistinguishable from latent representations obtained from the non-watermarked diffusion model. Therefore we can achieve watermark embedding with lossless performance for which we also provide theoretical proof. Furthermore since the watermark is intricately linked with image semantics it exhibits resilience to lossy processing and erasure attempts. The watermark can be extracted by Denoising Diffusion Implicit Models (DDIM) inversion and inverse sampling. We evaluate Gaussian Shading on multiple versions of Stable Diffusion and the results demonstrate that Gaussian Shading not only is performance-lossless but also outperforms existing methods in terms of robustness.
****
Generalized Event Cameras
Varun Sundar, Matthew Dutson, Andrei Ardelean, Claudio Bruschini, Edoardo Charbon, Mohit Gupta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25007-25017
Event cameras capture the world at high time resolution and with minimal bandwidth requirements. However event streams which only encode changes in brightness do not contain sufficient scene information to support a wide variety of downstream tasks. In this work we design generalized event cameras that inherently preserve scene intensity in a bandwidth-efficient manner. We generalize event cameras in terms of when an event is generated and what information is transmitted. To implement our designs we turn to single-photon sensors that provide digital access to individual photon detections; this modality gives us the flexibility to realize a rich space of generalized event cameras. Our single-photon event cameras are capable of high-speed high-fidelity imaging at low readout rates. Consequently these event cameras can support plug-and-play downstream inference without capturing new event datasets or designing specialized event-vision models. As a practical implication our designs which involve lightweight and near-sensor-compatible computations provide a way to use single-photon sensors without exorbitant bandwidth costs.
****
3D Neural Edge Reconstruction
Lei Li, Songyou Peng, Zehao Yu, Shaohui Liu, Rémi Pautrat, Xiaochuan Yin, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21219-21229
Real-world objects and environments are predominantly composed of edge features including straight lines and curves. Such edges are crucial elements for various applications such as CAD modeling surface meshing lane mapping etc. However existing traditional methods only prioritize lines over curves for simplicity in geometric modeling. To this end we introduce EMAP a new method for learning 3D edge representations with a focus on both lines and curves. Our method implicitly encodes 3D edge distance and direction in Unsigned Distance Functions (UDF) from multi-view edge maps. On top of this neural representation we propose an edge extraction algorithm that robustly abstracts parametric 3D edges from the inferred edge points and their directions. Comprehensive evaluations demonstrate that our method achieves better 3D edge reconstruction on multiple challenging datasets. We further show that our learned UDF field enhances neural surface reconstruction by capturing more details.
****
DocRes: A Generalist Model Toward Unifying Document Image Restoration Tasks
Jiaxin Zhang, Dezhi Peng, Chongyu Liu, Peirong Zhang, Lianwen Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15654-15664
Document image restoration is a crucial aspect of Document AI systems as the quality of document images significantly influences the overall performance. Prevailing methods address distinct restoration tasks independently leading to intricate systems and the incapability to harness the potential synergies of multi-task

learning. To overcome this challenge we propose DocRes a generalist model that unifies five document image restoration tasks including dewarping deshadowing appearance enhancement deblurring and binarization. To instruct DocRes to perform various restoration tasks we propose a novel visual prompt approach called Dynamic Task-Specific Prompt (DTSPrompt). The DTSPrompt for different tasks comprises distinct prior features which are additional characteristics extracted from the input image. Beyond its role as a cue for task-specific execution DTSPrompt can also serve as supplementary information to enhance the model's performance. Moreover DTSPrompt is more flexible than prior visual prompt approaches as it can be seamlessly applied and adapted to inputs with high and variable resolutions. Experimental results demonstrate that DocRes achieves competitive or superior performance compared to existing state-of-the-art task-specific models. This underscores the potential of DocRes across a broader spectrum of document image restoration tasks. The source code is publicly available at https://github.com/ZZZHANG-jx/DocRes.
****
Honeybee: Locality-enhanced Projector for Multimodal LLM
Junbum Cha, Wooyoung Kang, Jonghwan Mun, Byungseok Roh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13817-13827
In Multimodal Large Language Models (MLLMs) a visual projector plays a crucial role in bridging pre-trained vision encoders with LLMs enabling profound visual understanding while harnessing the LLMs' robust capabilities. Despite the importance of the visual projector it has been relatively less explored. In this study we first identify two essential projector properties: (i) flexibility in managing the number of visual tokens crucial for MLLMs' overall efficiency and (ii) preservation of local context from visual features vital for spatial understanding. Based on these findings we propose a novel projector design that is both flexible and locality-enhanced effectively satisfying the two desirable properties. Additionally we present comprehensive strategies to effectively utilize multiple and multifaceted instruction datasets. Through extensive experiments we examine the impact of individual design choices. Finally our proposed MLLM Honeybee remarkably outperforms previous state-of-the-art methods across various benchmarks including MME MMBench SEED-Bench and LLaVA-Bench achieving significantly higher efficiency. Code and models are available at https://github.com/kakaobrain/honeybee.
****
Learned Trajectory Embedding for Subspace Clustering
Yaroslava Lochman, Carl Olsson, Christopher Zach; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19092-19102
Clustering multiple motions from observed point trajectories is a fundamental task in understanding dynamic scenes. Most motion models require multiple tracks to estimate their parameters hence identifying clusters when multiple motions are observed is a very challenging task. This is even aggravated for high-dimensional motion models. The starting point of our work is that this high-dimensionality of motion model can actually be leveraged to our advantage as sufficiently long trajectories identify the underlying motion uniquely in practice. Consequently we propose to learn a mapping from trajectories to embedding vectors that represent the generating motion. The obtained trajectory embeddings are useful for clustering multiple observed motions but are also trained to contain sufficient information to recover the parameters of the underlying motion by utilizing a geometric loss. We therefore are able to use only weak supervision from given motion segmentation to train this mapping. The entire algorithm consisting of trajectory embedding clustering and motion parameter estimation is highly efficient. We conduct experiments on the Hopkins155 Hopkins12 and KT3DMoSeg datasets and show state-of-the-art performance of our proposed method for trajectory-based motion segmentation on full sequences and its competitiveness on the occluded sequences. Project page: https://ylochman.github.io/trajectory-embedding.
****

Training Vision Transformers for Semi-Supervised Semantic Segmentation
Xinting Hu, Li Jiang, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4007-4017
We present S4Former a novel approach to training Vision Transformers for Semi-Supervised Semantic Segmentation (S4). At its core S4Former employs a Vision Transformer within a classic teacher-student framework and then leverages three novel technical ingredients: PatchShuffle as a parameter-free perturbation technique Patch-Adaptive Self-Attention (PASA) as a fine-grained feature modulation method and the innovative Negative Class Ranking (NCR) regularization loss. Based on these regularization modules aligned with Transformer-specific characteristics across the image input feature and output dimensions S4Former exploits the Transformer's ability to capture and differentiate consistent global contextual information in unlabeled images. Overall S4Former not only defines a new state of the art in S4 but also maintains a streamlined and scalable architecture. Being readily compatible with existing frameworks S4Former achieves strong improvements (up to 4.9%) on benchmarks like Pascal VOC 2012 COCO and Cityscapes with varying numbers of labeled data. The code is at https://github.com/JoyHuYY1412/S4Former.
****
HarmonyView: Harmonizing Consistency and Diversity in One-Image-to-3D
Sangmin Woo, Byeongjun Park, Hyojun Go, Jin-Young Kim, Changick Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10574-10584
Recent progress in single-image 3D generation highlights the importance of multi-view coherency leveraging 3D priors from large-scale diffusion models pretrained on Internet-scale images. However the aspect of novel-view diversity remains underexplored within the research landscape due to the ambiguity in converting a 2D image into 3D content where numerous potential shapes can emerge. Here we aim to address this research gap by simultaneously addressing both consistency and diversity. Yet striking a balance between these two aspects poses a considerable challenge due to their inherent trade-offs. This work introduces HarmonyView a simple yet effective diffusion sampling technique adept at decomposing two intricate aspects in single-image 3D generation: consistency and diversity. This approach paves the way for a more nuanced exploration of the two critical dimensions within the sampling process. Moreover we propose a new evaluation metric based on CLIP image and text encoders to comprehensively assess the diversity of the generated views which closely aligns with human evaluators' judgments. In experiments HarmonyView achieves a harmonious balance demonstrating a win-win scenario in both consistency and diversity.
****
DGC-GNN: Leveraging Geometry and Color Cues for Visual Descriptor-Free 2D-3D Matching
Shuzhe Wang, Juho Kannala, Daniel Barath; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20881-20891
Matching 2D keypoints in an image to a sparse 3D point cloud of the scene without requiring visual descriptors has garnered increased interest due to its low memory requirements inherent privacy preservation and reduced need for expensive 3D model maintenance compared to visual descriptor-based methods. However existing algorithms often compromise on performance resulting in a significant deterioration compared to their descriptor-based counterparts. In this paper we introduce DGC-GNN a novel algorithm that employs a global-to-local Graph Neural Network (GNN) that progressively exploits geometric and color cues to rep- resent keypoints thereby improving matching accuracy. Our procedure encodes both Euclidean and angular relations at a coarse level forming the geometric embedding to guide the point matching. We evaluate DGC-GNN on both indoor and outdoor datasets demonstrating that it not only doubles the accuracy of the state-of-the-art visual descriptor-free algorithm but also substantially narrows the performance gap between descriptor-based and descriptor free methods.
****
CuVLER: Enhanced Unsupervised Object Discoveries through Exhaustive Self-Supervised Transformers

Shahaf Arica, Or Rubin, Sapir Gershov, Shlomi Laufer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23105-23114
In this paper we introduce VoteCut an innovative method for unsupervised object discovery that leverages feature representations from multiple self-supervised models. VoteCut employs normalized-cut based graph partitioning clustering and a pixel voting approach. Additionally We present CuVLER (Cut-Vote-and-LEaRn) a zero-shot model trained using pseudo-labels generated by VoteCut and a novel soft target loss to refine segmentation accuracy. Through rigorous evaluations across multiple datasets and several unsupervised setups our methods demonstrate significant improvements in comparison to previous state-of-the-art models. Our ablation studies further highlight the contributions of each component revealing the robustness and efficacy of our approach. Collectively VoteCut and CuVLER pave the way for future advancements in image segmentation.
****
Quantifying Task Priority for Multi-Task Optimization
Wooseong Jeong, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 363-372
The goal of multi-task learning is to learn diverse tasks within a single unified network. As each task has its own unique objective function conflicts emerge during training resulting in negative transfer among them. Earlier research identified these conflicting gradients in shared parameters between tasks and attempted to realign them in the same direction. However we prove that such optimization strategies lead to sub-optimal Pareto solutions due to their inability to accurately determine the individual contributions of each parameter across various tasks. In this paper we propose the concept of task priority to evaluate parameter contributions across different tasks. To learn task priority we identify the type of connections related to links between parameters influenced by task-specific losses during backpropagation. The strength of connections is gauged by the magnitude of parameters to determine task priority. Based on these we present a new method named connection strength-based optimization for multi-task learning which consists of two phases. The first phase learns the task priority within the network while the second phase modifies the gradients while upholding this priority. This ultimately leads to finding new Pareto optimal solutions for multiple tasks. Through extensive experiments we show that our approach greatly enhances multi-task performance in comparison to earlier gradient manipulation methods.
****
UnSAMFlow: Unsupervised Optical Flow Guided by Segment Anything Model
Shuai Yuan, Lei Luo, Zhuo Hui, Can Pu, Xiaoyu Xiang, Rakesh Ranjan, Denis Demandolx; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19027-19037
Traditional unsupervised optical flow methods are vulnerable to occlusions and motion boundaries due to lack of object-level information. Therefore we propose UnSAMFlow an unsupervised flow network that also leverages object information from the latest foundation model Segment Anything Model (SAM). We first include a self-supervised semantic augmentation module tailored to SAM masks. We also analyze the poor gradient landscapes of traditional smoothness losses and propose a new smoothness definition based on homography instead. A simple yet effective mask feature module has also been added to further aggregate features on the object level. With all these adaptations our method produces clear optical flow estimation with sharp boundaries around objects which outperforms state-of-the-art methods on both KITTI and Sintel datasets. Our method also generalizes well across domains and runs very efficiently.
****
Exploiting Inter-sample and Inter-feature Relations in Dataset Distillation
Wenxiao Deng, Wenbin Li, Tianyu Ding, Lei Wang, Hongguang Zhang, Kuihua Huang, Jing Huo, Yang Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17057-17066
Dataset distillation has emerged as a promising approach in deep learning enabling efficient training with small synthetic datasets derived from larger real one

s. Particularly distribution matching-based distillation methods attract attenti on thanks to its effectiveness and low computational cost. However these methods face two primary limitations: the dispersed feature distribution within the sam e class in synthetic datasets reducing class discrimination and an exclusive foc us on mean feature consistency lacking precision and comprehensiveness. To addre ss these challenges we introduce two novel constraints: a class centralization c onstraint and a covariance matching constraint. The class centralization constra int aims to enhance class discrimination by more closely clustering samples with in classes. The covariance matching constraint seeks to achieve more accurate fe ature distribution matching between real and synthetic datasets through local fe ature covariance matrices particularly beneficial when sample sizes are much sma ller than the number of features. Experiments demonstrate notable improvements w ith these constraints yielding performance boosts of up to 6.6% on CIFAR10 2.9% on SVHN 2.5% on CIFAR100 and 2.5% on TinyImageNet compared to the state-of-the-a rt relevant methods. In addition our method maintains robust performance in cros s-architecture settings with a maximum performance drop of 1.7% on four architec tures. Code is available at https://github.com/VincenDen/IID.
****

On the Scalability of Diffusion-based Text-to-Image Generation
Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R. Manmatha, Ashwin S waminathan, Zhuowen Tu, Stefano Ermon, Stefano Soatto; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9400-9409
Scaling up model and data size has been quite successful for the evolution of LL Ms. However the scaling law for the diffusion based text-to-image (T2I) models i s not fully explored. It is also unclear how to efficiently scale the model for better performance at reduced cost. The different training settings and expensiv e training cost make a fair model comparison extremely difficult. In this work w e empirically study the scaling properties of diffusion based T2I models by perf orming extensive and rigours ablations on scaling both denoising backbones and t raining set including training scaled UNet and Transformer variants ranging from 0.4B to 4B parameters on datasets upto 600M images. For model scaling we find t he location and amount of cross attention distinguishes the performance of exist ing UNet designs. And increasing the transformer blocks is more parameter-effici ent for improving text-image alignment than increasing channel numbers. We then identify an efficient UNet variant which is 45% smaller and 28% faster than SDXL 's UNet. On the data scaling side we show the quality and diversity of the train ing set matters more than simply dataset size. Increasing caption density and di versity improves text-image alignment performance and the learning efficiency. F inally we provide scaling functions to predict the text-image alignment performa nce as functions of the scale of model size compute and dataset size.
****

Entity-NeRF: Detecting and Removing Moving Entities in Urban Scenes
Takashi Otonari, Satoshi Ikehata, Kiyoharu Aizawa; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20892-209 01
Recent advancements in the study of Neural Radiance Fields (NeRF) for dynamic sc enes often involve explicit modeling of scene dynamics. However this approach fa ces challenges in modeling scene dynamics in urban environments where moving obj ects of various categories and scales are present. In such settings it becomes c rucial to effectively eliminate moving objects to accurately reconstruct static backgrounds. Our research introduces an innovative method termed here as Entity-NeRF which combines the strengths of knowledge-based and statistical strategies. This approach utilizes entity-wise statistics leveraging entity segmentation an d stationary entity classification through thing/stuff segmentation. To assess o ur methodology we created an urban scene dataset masked with moving objects. Our comprehensive experiments demonstrate that Entity-NeRF notably outperforms exis ting techniques in removing moving objects and reconstructing static urban backg rounds both quantitatively and qualitatively.
****

TAMM: TriAdapter Multi-Modal Learning for 3D Shape Understanding
Zhihao Zhang, Shengcao Cao, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21413-21423
The limited scale of current 3D shape datasets hinders the advancements in 3D shape understanding and motivates multi-modal learning approaches which transfer learned knowledge from data-abundant 2D image and language modalities to 3D shapes. However even though the image and language representations have been aligned by cross-modal models like CLIP we find that the image modality fails to contribute as much as the language in existing multi-modal 3D representation learning methods. This is attributed to the domain shift in the 2D images and the distinct focus of each modality. To more effectively leverage both modalities in the pre-training we introduce TriAdapter Multi-Modal Learning (TAMM) - a novel two-stage learning approach based on three synergistic adapters. First our CLIP Image Adapter mitigates the domain gap between 3D-rendered images and natural images by adapting the visual representations of CLIP for synthetic image-text pairs. Subsequently our Dual Adapters decouple the 3D shape representation space into two complementary sub-spaces: one focusing on visual attributes and the other for semantic understanding which ensure a more comprehensive and effective multi-modal pre-training. Extensive experiments demonstrate that TAMM consistently enhances 3D representations for a wide range of 3D encoder architectures pre-training datasets and downstream tasks. Notably we boost the zero-shot classification accuracy on Objaverse-LVIS from 46.8% to 50.7% and improve the 5-way 10-shot linear probing classification accuracy on ModelNet40 from 96.1% to 99.0%. Project page: https://alanzhangcs.github.io/tamm-page.
****
GauHuman: Articulated Gaussian Splatting from Monocular Human Videos
Shoukang Hu, Tao Hu, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20418-20431
We present GauHuman a 3D human model with Gaussian Splatting for both fast training (1 2 minutes) and real-time rendering (up to 189 FPS) compared with existing NeRF-based implicit representation modelling frameworks demanding hours of training and seconds of rendering per frame. Specifically GauHuman encodes Gaussian Splatting in the canonical space and transforms 3D Gaussians from canonical space to posed space with linear blend skinning (LBS) in which effective pose and LBS refinement modules are designed to learn fine details of 3D humans under negligible computational cost. Moreover to enable fast optimization of GauHuman we initialize and prune 3D Gaussians with 3D human prior while splitting/cloning via KL divergence guidance along with a novel merge operation for further speeding up. Extensive experiments on ZJU_Mocap and MonoCap datasets demonstrate that GauHuman achieves state-of-the-art performance quantitatively and qualitatively with fast training and real-time rendering speed. Notably without sacrificing rendering quality GauHuman can fast model the 3D human performer with 13k 3D Gaussians. Our code is available at https://github.com/skhu101/GauHuman.
****
AnySkill: Learning Open-Vocabulary Physical Skill for Interactive Agents
Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, Siyuan Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 852-862
Traditional approaches in physics-based motion generation centered around imitation learning and reward shaping often struggle to adapt to new scenarios. To tackle this limitation we propose AnySkill a novel hierarchical method that learns physically plausible interactions following open-vocabulary instructions. Our approach begins by developing a set of atomic actions via a low-level controller trained via imitation learning. Upon receiving an open-vocabulary textual instruction AnySkill employs a high-level policy that selects and integrates these atomic actions to maximize the CLIP similarity between the agent's rendered images and the text. An important feature of our method is the use of image-based rewards for the high-level policy which allows the agent to learn interactions with objects without manual reward engineering. We demonstrate AnySkill's capability to generate realistic and natural motion sequences in response to unseen instructi

ons of varying lengths marking it the first method capable of open-vocabulary ph
ysical skill learning for interactive humanoid agents.
****
EGTR: Extracting Graph from Transformer for Scene Graph Generation
Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, Seunghyun Park; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 24229-24238
Scene Graph Generation (SGG) is a challenging task of detecting objects and pred
icting relationships between objects. After DETR was developed one-stage SGG mod
els based on a one-stage object detector have been actively studied. However com
plex modeling is used to predict the relationship between objects and the inhere
nt relationship between object queries learned in the multi-head self-attention
of the object detector has been neglected. We propose a lightweight one-stage SG
G model that extracts the relation graph from the various relationships learned
in the multi-head self-attention layers of the DETR decoder. By fully utilizing
the self-attention by-products the relation graph can be extracted effectively w
ith a shallow relation extraction head. Considering the dependency of the relati
on extraction task on the object detection task we propose a novel relation smoo
thing technique that adjusts the relation label adaptively according to the qual
ity of the detected objects. By the relation smoothing the model is trained acco
rding to the continuous curriculum that focuses on object detection task at the
beginning of training and performs multi-task learning as the object detection p
erformance gradually improves. Furthermore we propose a connectivity prediction
task that predicts whether a relation exists between object pairs as an auxiliar
y task of the relation extraction. We demonstrate the effectiveness and efficien
cy of our method for the Visual Genome and Open Image V6 datasets. Our code is p
ublicly available at https://github.com/naver-ai/egtr.
****
Generative Unlearning for Any Identity
Juwon Seo, Sung-Hoon Lee, Tae-Young Lee, Seungjun Moon, Gyeong-Moon Park; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2024, pp. 9151-9161
Recent advances in generative models trained on large-scale datasets have made i
t possible to synthesize high-quality samples across various domains. Moreover t
he emergence of strong inversion networks enables not only a reconstruction of r
eal-world images but also the modification of attributes through various editing
 methods. However in certain domains related to privacy issues e.g. human faces
advanced generative models along with strong inversion methods can lead to poten
tial misuses. In this paper we propose an essential yet under-explored task call
ed generative identity unlearning which steers the model not to generate an imag
e of a specific identity. In the generative identity unlearning we target the fo
llowing objectives: (i) preventing the generation of images with a certain ident
ity and (ii) preserving the overall quality of the generative model. To satisfy
these goals we propose a novel framework Generative Unlearning for Any Identity
(GUIDE) which prevents the reconstruction of a specific identity by unlearning t
he generator with only a single image. GUIDE consists of two parts: (i) finding
a target point for optimization that un-identifies the source latent code and (i
i) novel loss functions that facilitate the unlearning procedure while less affe
cting the learned distribution. Our extensive experiments demonstrate that our p
roposed method achieves state-of-the-art performance in the generative machine u
nlearning task. The code is available at https://github.com/KHU-AGI/GUIDE.
****
Context-based and Diversity-driven Specificity in Compositional Zero-Shot Learni
ng
Yun Li, Zhe Liu, Hang Chen, Lina Yao; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17037-17046
Compositional Zero-Shot Learning (CZSL) aims to recognize unseen attribute-objec
t pairs based on a limited set of observed examples. Current CZSL methodologies
despite their advancements tend to neglect the distinct specificity levels prese
nt in attributes. For instance given images of sliced strawberries they may fail

to prioritize `Sliced-Strawberry' over a generic `Red-Strawberry' despite the former being more informative. They also suffer from ballooning search space when shifting from Close-World (CW) to Open-World (OW) CZSL. To address the issues we introduce the Context-based and Diversity-driven Specificity learning framework for CZSL (CDS-CZSL). Our framework evaluates the specificity of attributes by considering the diversity of objects they apply to and their related context. This novel approach allows for more accurate predictions by emphasizing specific attribute-object pairs and improves composition filtering in OW-CZSL. We conduct experiments in both CW and OW scenarios and our model achieves state-of-the-art results across three datasets.
****
FlowVid: Taming Imperfect Optical Flows for Consistent Video-to-Video Synthesis
Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, Diana Marculescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8207-8216
Diffusion models have transformed the image-to-image (I2I) synthesis and are now permeating into videos. However the advancement of video-to-video (V2V) synthesis has been hampered by the challenge of maintaining temporal consistency across video frames. This paper proposes a consistent V2V synthesis framework by jointly leveraging spatial conditions and temporal optical flow clues within the source video. Contrary to prior methods that strictly adhere to optical flow our approach harnesses its benefits while handling the imperfection in flow estimation. We encode the optical flow via warping from the first frame and serve it as a supplementary reference in the diffusion model. This enables our model for video synthesis by editing the first frame with any prevalent I2I models and then propagating edits to successive frames. Our V2V model FlowVid demonstrates remarkable properties: (1) Flexibility: FlowVid works seamlessly with existing I2I models facilitating various modifications including stylization object swaps and local edits. (2) Efficiency: Generation of a 4-second video with 30 FPS and 512x512 resolution takes only 1.5 minutes which is 3.1x 7.2x and 10.5x faster than CoDeF Rerender and TokenFlow respectively. (3) High-quality: In user studies our FlowVid is preferred 45.7% of the time outperforming CoDeF (3.5%) Rerender (10.2%) and TokenFlow (40.4%).
****
StyleCineGAN: Landscape Cinemagraph Generation using a Pre-trained StyleGAN
Jongwoo Choi, Kwanggyoon Seo, Amirsaman Ashtari, Junyong Noh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7872-7881
We propose a method that can generate cinemagraphs automatically from a still landscape image using a pre-trained StyleGAN. Inspired by the success of recent unconditional video generation we leverage a powerful pre-trained image generator to synthesize high-quality cinemagraphs. Unlike previous approaches that mainly utilize the latent space of a pre-trained StyleGAN our approach utilizes its deep feature space for both GAN inversion and cinemagraph generation. Specifically we propose multi-scale deep feature warping (MSDFW) which warps the intermediate features of a pre-trained StyleGAN at different resolutions. By using MSDFW the generated cinemagraphs are of high resolution and exhibit plausible looping animation. We demonstrate the superiority of our method through user studies and quantitative comparisons with state-of-the-art cinemagraph generation methods and a video generation method that uses a pre-trained StyleGAN.
****
Rethinking Multi-domain Generalization with A General Learning Objective
Zhaorui Tan, Xi Yang, Kaizhu Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23512-23522
Multi-domain generalization (mDG) is universally aimed to minimize the discrepancy between training and testing distributions to enhance marginal-to-label distribution mapping. However existing mDG literature lacks a general learning objective paradigm and often imposes constraints on static target marginal distributions. In this paper we propose to leverage a Y-mapping to relax the constraint. We

rethink the learning objective for mDG and design a new general learning object
ive to interpret and analyze most existing mDG wisdom. This general objective is
 bifurcated into two synergistic amis: learning domain-independent conditional f
eatures and maximizing a posterior. Explorations also extend to two effective re
gularization terms that incorporate prior information and suppress invalid causa
lity alleviating the issues that come with relaxed constraints. We theoretically
 contribute an upper bound for the domain alignment of domain-independent condit
ional features disclosing that many previous mDG endeavors actually optimize par
tially the objective and thus lead to limited performance. As such our study dis
tills a general learning objective into four practical components providing a ge
neral robust and flexible mechanism to handle complex domain shifts. Extensive e
mpirical results indicate that the proposed objective with Y-mapping leads to su
bstantially better mDG performance in various downstream tasks including regress
ion segmentation and classification. Code is available at https://github.com/zha
orui-tan/GMDG/tree/main.
****
Laplacian-guided Entropy Model in Neural Codec with Blur-dissipated Synthesis
Atefeh Khoshkhahtinat, Ali Zafari, Piyush M. Mehta, Nasser M. Nasrabadi; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 3045-3054
While replacing Gaussian decoders with a conditional diffusion model enhances th
e perceptual quality of reconstructions in neural image compression their lack o
f inductive bias for image data restricts their ability to achieve state-of-the-
art perceptual levels. To address this limitation we adopt a non-isotropic diffu
sion model at the decoder side. This model imposes an inductive bias aimed at di
stinguishing between frequency contents thereby facilitating the generation of h
igh-quality images. Moreover our framework is equipped with a novel entropy mode
l that accurately models the probability distribution of latent representation b
y exploiting spatio-channel correlations in latent space while accelerating the
entropy decoding step. This channel-wise entropy model leverages both local and
global spatial contexts within each channel chunk. The global spatial context is
 built upon the Transformer which is specifically designed for image compression
 tasks. The designed Transformer employs a Laplacian-shaped positional encoding
the learnable parameters of which are adaptively adjusted for each channel clust
er. Our experiments demonstrate that our proposed framework yields better percep
tual quality compared to cutting-edge generative-based codecs and the proposed e
ntropy model contributes to notable bitrate savings. The code is available at ht
tps://github.com/Atefeh-Khoshtinat/Blur-dissipated-compression.
****
Universal Novelty Detection Through Adaptive Contrastive Learning
Hossein Mirzaei, Mojtaba Nafez, Mohammad Jafari, Mohammad Bagher Soltani, Mohamm
ad Azizmalayeri, Jafar Habibi, Mohammad Sabokrou, Mohammad Hossein Rohban; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 22914-22923
Novelty detection is a critical task for deploying machine learning models in th
e open world. A crucial property of novelty detection methods is universality wh
ich can be interpreted as generalization across various distributions of trainin
g or test data. More precisely for novelty detection distribution shifts may occ
ur in the training set or the test set. Shifts in the training set refer to case
s where we train a novelty detector on a new dataset and expect strong transfera
bility. Conversely distribution shifts in the test set indicate the methods' per
formance when the trained model encounters a shifted test sample. We experimenta
lly show that existing methods falter in maintaining universality which stems fr
om their rigid inductive biases. Motivated by this we aim for more generalized t
echniques that have more adaptable inductive biases. In this context we leverage
 the fact that contrastive learning provides an efficient framework to easily sw
itch and adapt to new inductive biases through the proper choice of augmentation
s in forming the negative pairs. We propose a novel probabilistic auto-negative
pair generation method AutoAugOOD along with contrastive learning to yield a uni
versal novelty detector method. Our experiments demonstrate the superiority of o

ur method under different distribution shifts in various image benchmark datasets. Notably our method emerges universality in the lens of adaptability to different setups of novelty detection including one-class unlabeled multi-class and labeled multi-class settings.
****

Rethinking Diffusion Model for Multi-Contrast MRI Super-Resolution
Guangyuan Li, Chen Rao, Juncheng Mo, Zhanjie Zhang, Wei Xing, Lei Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11365-11374
Recently diffusion models (DM) have been applied in magnetic resonance imaging (MRI) super-resolution (SR) reconstruction exhibiting impressive performance especially with regard to detailed reconstruction. However the current DM-based SR reconstruction methods still face the following issues: (1) They require a large number of iterations to reconstruct the final image which is inefficient and consumes a significant amount of computational resources. (2) The results reconstructed by these methods are often misaligned with the real high-resolution images leading to remarkable distortion in the reconstructed MR images. To address the aforementioned issues we propose an efficient diffusion model for multi-contrast MRI SR named as DiffMSR. Specifically we apply DM in a highly compact low-dimensional latent space to generate prior knowledge with high-frequency detail information. The highly compact latent space ensures that DM requires only a few simple iterations to produce accurate prior knowledge. In addition we design the Prior-Guide Large Window Transformer (PLWformer) as the decoder for DM which can extend the receptive field while fully utilizing the prior knowledge generated by DM to ensure that the reconstructed MR image remains undistorted. Extensive experiments on public and clinical datasets demonstrate that our DiffMSR outperforms state-of-the-art methods.
****

Resurrecting Old Classes with New Data for Exemplar-Free Continual Learning
Dipam Goswami, Albin Soutif-Cormerais, Yuyang Liu, Sandesh Kamath, Bart?omiej Twardowski, Joost van de Weijer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28525-28534
Continual learning methods are known to suffer from catastrophic forgetting a phenomenon that is particularly hard to counter for methods that do not store exemplars of previous tasks. Therefore to reduce potential drift in the feature extractor existing exemplar-free methods are typically evaluated in settings where the first task is significantly larger than subsequent tasks. Their performance drops drastically in more challenging settings starting with a smaller first task. To address this problem of feature drift estimation for exemplar-free methods we propose to adversarially perturb the current samples such that their embeddings are close to the old class prototypes in the old model embedding space. We then estimate the drift in the embedding space from the old to the new model using the perturbed images and compensate the prototypes accordingly. We exploit the fact that adversarial samples are transferable from the old to the new feature space in a continual learning setting. The generation of these images is simple and computationally cheap. We demonstrate in our experiments that the proposed approach better tracks the movement of prototypes in embedding space and outperforms existing methods on several standard continual learning benchmarks as well as on fine-grained datasets. Code is available at https://github.com/dipamgoswami/ADC.
****

Unknown Prompt the only Lacuna: Unveiling CLIP's Potential for Open Domain Generalization
Mainak Singha, Ankit Jha, Shirsha Bose, Ashwin Nair, Moloud Abdar, Biplab Banerjee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13309-13319
We delve into Open Domain Generalization (ODG) marked by domain and category shifts between training's labeled source and testing's unlabeled target domains. Existing solutions to ODG face limitations due to constrained generalizations of traditional CNN backbones and errors in detecting target open samples in the abse

nce of prior knowledge. Addressing these pitfalls we introduce ODG-CLIP harnessi
ng the semantic prowess of the vision-language model CLIP. Our framework brings
forth three primary innovations: Firstly distinct from prevailing paradigms we c
onceptualize ODG as a multi-class classification challenge encompassing both kno
wn and novel categories. Central to our approach is modeling a unique prompt tai
lored for detecting unknown class samples and to train this we employ a readily
accessible stable diffusion model elegantly generating proxy images for the open
 class. Secondly aiming for domain-tailored classification (prompt) weights whil
e ensuring a balance of precision and simplicity we devise a novel visual style-
centric prompt learning mechanism. Finally we infuse images with class-discrimin
ative knowledge derived from the prompt space to augment the fidelity of CLIP's
visual embeddings. We introduce a novel objective to safeguard the continuity of
 this infused semantic intel across domains especially for the shared classes. T
hrough rigorous testing on diverse datasets covering closed and open-set DG cont
exts ODG-CLIP demonstrates clear supremacy consistently outpacing peers with per
formance boosts between 8%-16%. Code will be available at https://github.com/mai
naksingha01/ODG-CLIP.
****
Poly Kernel Inception Network for Remote Sensing Detection
Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, Yazhou Yao; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 27706-27716
Object detection in remote sensing images (RSIs) often suffers from several incr
easing challenges including the large variation in object scales and the diverse
-ranging context. Prior methods tried to address these challenges by expanding t
he spatial receptive field of the backbone either through large-kernel convoluti
on or dilated convolution. However the former typically introduces considerable
background noise while the latter risks generating overly sparse feature represe
ntations. In this paper we introduce the Poly Kernel Inception Network (PKINet)
to handle the above challenges. PKINet employs multi-scale convolution kernels w
ithout dilation to extract object features of varying scales and capture local c
ontext. In addition a Context Anchor Attention (CAA) module is introduced in par
allel to capture long-range contextual information. These two components work jo
intly to advance the performance of PKINet on four challenging remote sensing ob
ject detection benchmarks namely DOTA-v1.0 DOTA-v1.5 HRSC2016 and DIOR-R.
****
RMT: Retentive Networks Meet Vision Transformers
Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, Ran He; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 5641-5651
Vision Transformer (ViT) has gained increasing attention in the computer vision
community in recent years. However the core component of ViT Self-Attention lack
s explicit spatial priors and bears a quadratic computational complexity thereby
 constraining the applicability of ViT. To alleviate these issues we draw inspir
ation from the recent Retentive Network (RetNet) in the field of NLP and propose
 RMT a strong vision backbone with explicit spatial prior for general purposes.
Specifically we extend the RetNet's temporal decay mechanism to the spatial doma
in and propose a spatial decay matrix based on the Manhattan distance to introdu
ce the explicit spatial prior to Self-Attention. Additionally an attention decom
position form that adeptly adapts to explicit spatial prior is proposed aiming t
o reduce the computational burden of modeling global information without disrupt
ing the spatial decay matrix. Based on the spatial decay matrix and the attentio
n decomposition form we can flexibly integrate explicit spatial prior into the v
ision backbone with linear complexity. Extensive experiments demonstrate that RM
T exhibits exceptional performance across various vision tasks. Specifically wit
hout extra training data RMT achieves 84.8% and 86.1% top-1 acc on ImageNet-1k w
ith 27M/4.5GFLOPs and 96M/18.2GFLOPs. For downstream tasks RMT achieves 54.5 box
 AP and 47.2 mask AP on the COCO detection task and 52.8 mIoU on the ADE20K sema
ntic segmentation task.
****

From Coarse to Fine-Grained Open-Set Recognition
Nico Lang, Vésteinn Snæbjarnarson, Elijah Cole, Oisin Mac Aodha, Christian Igel, Serge Belongie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17804-17814
Open-set recognition (OSR) methods aim to identify whether or not a test example belongs to a category ob- served during training. Depending on how visually sim - ilar a test example is to the training categories the OSR task can be easy or extremely challenging. However the vast majority of previous work has studied OSR in the presence of large coarse-grained semantic shifts. In contrast many real-world problems are inherently fine- grained which means that test examples may be highly visually similar to the training categories. Motivated by this observation we investigate three aspects of OSR: label granularity similarity between the open- and closed-sets and the role of hierarchical supervision during training. To study these dimensions we curate new open-set splits of a large fine-grained visual categorization dataset. Our anal- ysis results in several interesting findings including: (i) the best OSR method to use is heavily dependent on the degree of semantic shift present and (ii) hierarchical rep- resentation learning can improve coarse-grained OSR but has little effect on fine-grained OSR performance. To fur- ther enhance fine-grained OSR performance we propose a hierarchy-adversarial learning method to discourage hier- archical structure in the representation space which results in a perhaps counter-intuitive behaviour and a relative im- provement in fine-grained OSR of up to 2% in AUROC and 7% in AUPR over standard training. Code and data are available: langnico.github.io/fine-grained-osr.
****
Multimodal Pathway: Improve Transformers with Irrelevant Data from Other Modalities
Yiyuan Zhang, Xiaohan Ding, Kaixiong Gong, Yixiao Ge, Ying Shan, Xiangyu Yue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6108-6117
We propose to improve transformers of a specific modality with irrelevant data from other modalities e.g. improve an ImageNet model with audio or point cloud datasets. We would like to highlight that the data samples of the target modality are irrelevant to the other modalities which distinguishes our method from other works utilizing paired (e.g. CLIP) or interleaved data of different modalities. We propose a methodology named Multimodal Pathway - given a target modality and a transformer designed for it we use an auxiliary transformer trained with data of another modality and construct pathways to connect components of the two models so that data of the target modality can be processed by both models. In this way we utilize the universal sequence-to-sequence modeling abilities of transformers obtained from two modalities. As a concrete implementation we use a modality-specific tokenizer and task-specific head as usual but utilize the transformer blocks of the auxiliary model via a proposed method named Cross-Modal Re-parameterization which exploits the auxiliary weights without any inference costs. On the image point cloud video and audio recognition tasks we observe significant and consistent performance improvements with irrelevant data from other modalities. The code and models are available at https://github.com/AILab-CVC/M2PT.
****
FaceChain-ImagineID: Freely Crafting High-Fidelity Diverse Talking Faces from Disentangled Audio
Chao Xu, Yang Liu, Jiazheng Xing, Weida Wang, Mingze Sun, Jun Dan, Tianxin Huang, Siyuan Li, Zhi-Qi Cheng, Ying Tai, Baigui Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1292-1302
In this paper we abstract the process of people hearing speech extracting meaningful cues and creating various dynamically audio-consistent talking faces termed Listening and Imagining into the task of high-fidelity diverse talking faces generation from a single audio. Specifically it involves two critical challenges: one is to effectively decouple identity content and emotion from entangled audio and the other is to maintain intra-video diversity and inter-video consistency. To tackle the issues we first dig out the intricate relationships among facial

factors and simplify the decoupling process tailoring a Progressive Audio Disent anglement for accurate facial geometry and semantics learning where each stage i ncorporates a customized training module responsible for a specific factor. Seco ndly to achieve visually diverse and audio-synchronized animation solely from in put audio within a single model we introduce the Controllable Coherent Frame gen eration which involves the flexible integration of three trainable adapters with frozen Latent Diffusion Models (LDMs) to focus on maintaining facial geometry a nd semantics as well as texture and temporal coherence between frames. In this w ay we inherit high-quality diverse generation from LDMs while significantly impr oving their controllability at a low training cost. Extensive experiments demons trate the flexibility and effectiveness of our method in handling this paradigm. The codes will be released at https://github.com/modelscope/facechain.
****
OmniViD: A Generative Framework for Universal Video Understanding
Junke Wang, Dongdong Chen, Chong Luo, Bo He, Lu Yuan, Zuxuan Wu, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2024, pp. 18209-18220
The core of video understanding tasks such as recognition captioning and trackin g is to automatically detect objects or actions in a video and analyze their tem poral evolution. Despite sharing a common goal different tasks often rely on dis tinct model architectures and annotation formats. In contrast natural language p rocessing benefits from a unified output space i.e. text sequences which simplif ies the training of powerful foundational language models such as GPT-3 with ext ensive training corpora. Inspired by this we seek to unify the output space of v ideo understanding tasks by using languages as labels and additionally introduci ng time and box tokens. In this way a variety of video tasks could be formulated as video-grounded token generation. This enables us to address various types of video tasks including classification (such as action recognition) captioning (c overing clip captioning video question answering and dense video captioning) and localization tasks (such as visual object tracking) within a fully shared encod er-decoder architecture following a generative framework. Through comprehensive experiments we demonstrate such a simple and straightforward idea is quite effec tive and can achieve state-of-the-art or competitive results on seven video benc hmarks providing a novel perspective for more universal video understanding. Cod e is available at \href https://github.com/wangjk666/OmniVid https://github.com /wangjk666/OmniVid .
****
Naturally Supervised 3D Visual Grounding with Language-Regularized Concept Learn ers
Chun Feng, Joy Hsu, Weiyu Liu, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13269-13278
3D visual grounding is a challenging task that often requires direct and dense s upervision notably the semantic label for each object in the scene. In this pape r we instead study the naturally supervised setting that learns from only 3D sce ne and QA pairs where prior works underperform. We propose the Language-Regulari zed Concept Learner (LARC) which uses constraints from language as regularizatio n to significantly improve the accuracy of neuro-symbolic concept learners in th e naturally supervised setting. Our approach is based on two core insights: the first is that language constraints (e.g. a word's relation to another) can serve as effective regularization for structured representations in neuro-symbolic mo dels; the second is that we can query large language models to distill such cons traints from language properties. We show that LARC improves performance of prio r works in naturally supervised 3D visual grounding and demonstrates a wide rang e of 3D visual reasoning capabilities--from zero-shot composition to data effici ency and transferability. Our method represents a promising step towards regular izing structured visual reasoning frameworks with language-based priors for lear ning in settings without dense supervision.
****
SSR-Encoder: Encoding Selective Subject Representation for Subject-Driven Genera tion

Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, Zhongliang Jing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8069-8078
Recent advancements in subject-driven image generation have led to zero-shot generation yet precise selection and focus on crucial subject representations remain challenging. Addressing this we introduce the SSR-Encoder a novel architecture designed for selectively capturing any subject from single or multiple reference images. It responds to various query modalities including text and masks without necessitating test-time fine-tuning. The SSR-Encoder combines a Token-to-Patch Aligner that aligns query inputs with image patches and a Detail-Preserving Subject Encoder for extracting and preserving fine features of the subjects thereby generating subject embeddings. These embeddings used in conjunction with original text embeddings condition the generation process. Characterized by its model generalizability and efficiency the SSR-Encoder adapts to a range of custom models and control modules. Enhanced by the Embedding Consistency Regularization Loss for improved training our extensive experiments demonstrate its effectiveness in versatile and high-quality image generation indicating its broad applicability.
****
CA-Jaccard: Camera-aware Jaccard Distance for Person Re-identification
Yiyu Chen, Zheyi Fan, Zhaoru Chen, Yixuan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17532-17541
Person re-identification (re-ID) is a challenging task that aims to learn discriminative features for person retrieval. In person re-ID Jaccard distance is a widely used distance metric especially in re-ranking and clustering scenarios. However we discover that camera variation has a significant negative impact on the reliability of Jaccard distance. In particular Jaccard distance calculates the distance based on the overlap of relevant neighbors. Due to camera variation intra-camera samples dominate the relevant neighbors which reduces the reliability of the neighbors by introducing intra-camera negative samples and excluding inter-camera positive samples. To overcome this problem we propose a novel camera-aware Jaccard (CA-Jaccard) distance that leverages camera information to enhance the reliability of Jaccard distance. Specifically we design camera-aware k-reciprocal nearest neighbors (CKRNNs) to find k-reciprocal nearest neighbors on the intra-camera and inter-camera ranking lists which improves the reliability of relevant neighbors and guarantees the contribution of inter-camera samples in the overlap. Moreover we propose a camera-aware local query expansion (CLQE) to mine reliable samples in relevant neighbors by exploiting camera variation as a strong constraint and assign these samples higher weights in overlap further improving the reliability. Our CA-Jaccard distance is simple yet effective and can serve as a general distance metric for person re-ID methods with high reliability and low computational cost. Extensive experiments demonstrate the effectiveness of our method. Code is available at https://github.com/chen960/CA-Jaccard/.
****
Dual Prior Unfolding for Snapshot Compressive Imaging
Jiancheng Zhang, Haijin Zeng, Jiezhang Cao, Yongyong Chen, Dengxiu Yu, Yin-Ping Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25742-25752
Recently deep unfolding methods have achieved remarkable success in the realm of Snapshot Compressive Imaging (SCI) reconstruction. However the existing methods all follow the iterative framework of a single image prior which limits the efficiency of the unfolding methods and makes it a problem to use other priors simply and effectively. To break out of the box we derive an effective Dual Prior Unfolding (DPU) which achieves the joint utilization of multiple deep priors and greatly improves iteration efficiency. Our unfolding method is implemented through two parts i.e. Dual Prior Framework (DPF) and Focused Attention (FA). In brief in addition to the normal image prior DPF introduces a residual into the iteration formula and constructs a degraded prior for the residual by considering various degradations to establish the unfolding framework. To improve the effectiveness of the image prior based on self-attention FA adopts a novel mechanism inspi

red by PCA denoising to scale and filter attention which lets the attention focu s more on effective features with little computation cost. Besides an asymmetric backbone is proposed to further improve the efficiency of hierarchical self-att ention. Remarkably our 5-stage DPU achieves state-of-the-art (SOTA) performance with the least FLOPs and parameters compared to previous methods while our 9-sta ge DPU significantly outperforms other unfolding methods with less computational requirement.
****
COLMAP-Free 3D Gaussian Splatting
Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, Xiaolong Wang; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20796-20805
While neural rendering has led to impressive advances in scene reconstruction an d novel view synthesis it relies heavily on accurately pre-computed camera poses . To relax this constraint multiple efforts have been made to train Neural Radia nce Fields (NeRFs) without pre-processed camera poses. However the implicit repr esentations of NeRFs provide extra challenges to optimize the 3D structure and c amera poses at the same time. On the other hand the recently proposed 3D Gaussia n Splatting provides new opportunities given its explicit point cloud representa tions. This paper leverages both the explicit geometric representation and the c ontinuity of the input video stream to perform novel view synthesis without any SfM preprocessing. We process the input frames in a sequential manner and progre ssively grow the 3D Gaussians set by taking one input frame at a time without th e need to pre-compute the camera poses. Our method significantly improves over p revious approaches in view synthesis and camera pose estimation under large moti on changes. Our project page is: https: //oasisyang.github.io/colmap-free-3dgs.
****
MVIP-NeRF: Multi-view 3D Inpainting on NeRF Scenes via Diffusion Prior
Honghua Chen, Chen Change Loy, Xingang Pan; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5344-5353
Despite the emergence of successful NeRF inpainting methods built upon explicit RGB and depth 2D inpainting supervisions these methods are inherently constraine d by the capabilities of their underlying 2D inpainters. This is due to two key reasons: (i) independently inpainting constituent images results in view-inconsi stent imagery and (ii) 2D inpainters struggle to ensure high-quality geometry co mpletion and alignment with inpainted RGB images. To overcome these limitations we propose a novel approach called MVIP-NeRF that harnesses the potential of dif fusion priors for NeRF inpainting addressing both appearance and geometry aspect s. MVIP-NeRF performs joint inpainting across multiple views to reach a consiste nt solution which is achieved via an iterative optimization process based on Sco re Distillation Sampling (SDS). Apart from recovering the rendered RGB images we also extract normal maps as a geometric representation and define a normal SDS loss that motivates accurate geometry inpainting and alignment with the appearan ce. Additionally we formulate a multi-view SDS score function to distill generat ive priors simultaneously from different view images ensuring consistent visual completion when dealing with large view variations. Our experimental results sho w better appearance and geometry recovery than previous NeRF inpainting methods.
****
StegoGAN: Leveraging Steganography for Non-Bijective Image-to-Image Translation
Sidi Wu, Yizi Chen, Samuel Mermet, Lorenz Hurni, Konrad Schindler, Nicolas Gonth ier, Loic Landrieu; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2024, pp. 7922-7931
Most image-to-image translation models postulate that a unique correspondence ex ists between the semantic classes of the source and target domains. However this assumption does not always hold in real-world scenarios due to divergent distri butions different class sets and asymmetrical information representation. As con ventional GANs attempt to generate images that match the distribution of the tar get domain they may hallucinate spurious instances of classes absent from the so urce domain thereby diminishing the usefulness and reliability of translated ima ges. CycleGAN-based methods are also known to hide the mismatched information in

the generated images to bypass cycle consistency objectives a process known as steganography. In response to the challenge of non-bijective image translation we introduce StegoGAN a novel model that leverages steganography to prevent spurious features in generated images. Our approach enhances the semantic consistency of the translated images without requiring additional postprocessing or supervision. Our experimental evaluations demonstrate that StegoGAN outperforms existing GAN-based models across various non-bijective image-to-image translation tasks both qualitatively and quantitatively. Our code and pretrained models are accessible at https://github.com/sian-wusidi/StegoGAN.
****
M&M VTO: Multi-Garment Virtual Try-On and Editing
Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1346-1356
We present M&M VTO-a mix and match virtual try-on method that takes as input multiple garment images text description for garment layout and an image of a person. An example input includes: an image of a shirt an image of a pair of pants "rolled sleeves shirt tucked in" and an image of a person. The output is a visualization of how those garments (in the desired layout) would look like on the given person. Key contributions of our method are: 1) a single stage diffusion based model with no super resolution cascading that allows to mix and match multiple garments at 1024x512 resolution preserving and warping intricate garment details 2) architecture design (VTO UNet Diffusion Transformer) to disentangle denoising from person specific features allowing for a highly effective finetuning strategy for identity preservation (6MB model per individual vs 4GB achieved with e.g. dreambooth finetuning); solving a common identity loss problem in current virtual try-on methods 3) layout control for multiple garments via text inputs finetuned over PaLI-3 for virtual try-on task. Experimental results indicate that M&M VTO achieves state-of-the-art performance both qualitatively and quantitatively as well as opens up new opportunities for virtual try-on via language-guided and multi-garment try-on.
****
AutoAD III: The Prequel - Back to the Pixels
Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18164-18174
Generating Audio Description (AD) for movies is a challenging task that requires fine-grained visual understanding and an awareness of the characters and their names. Currently visual language models for AD generation are limited by a lack of suitable training data and also their evaluation is hampered by using performance measures not specialized to the AD domain. In this paper we make three contributions: (i) We propose two approaches for constructing AD datasets with aligned video data and build training and evaluation datasets using these. These datasets will be publicly released; (ii) We develop a Q-former-based architecture which ingests raw video and generates AD using frozen pre-trained visual encoders and large language models; and (iii) We provide new evaluation metrics to benchmark AD quality that are well matched to human performance. Taken together we improve the state of the art on AD generation.
****
Characteristics Matching Based Hash Codes Generation for Efficient Fine-grained Image Retrieval
Zhen-Duo Chen, Li-Jun Zhao, Zi-Chao Zhang, Xin Luo, Xin-Shun Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17273-17281
The rapidly growing scale of data in practice poses demands on the efficiency of retrieval models. However for fine-grained image retrieval task there are inherent contradictions in the design of hashing based efficient models. Firstly the limited information embedding capacity of low-dimensional binary hash codes coupled with the detailed information required to describe fine-grained categories results in a contradiction in feature learning. Secondly there is also a contradi

ction between the complexity of fine-grained feature extraction models and retri
eval efficiency. To address these issues in this paper we propose the characteri
stics matching based hash codes generation method. Coupled with the cross-layer
semantic information transfer module and the multi-region feature embedding modu
le the proposed method can generate hash codes that effectively capture fine-gra
ined differences among samples while ensuring efficient inference. Extensive exp
eriments on widely used datasets demonstrate that our method can significantly o
utperform state-of-the-art methods.
****
BadCLIP: Dual-Embedding Guided Backdoor Attack on Multimodal Contrastive Learnin
g
Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, Ee-Chien Chang;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2024, pp. 24645-24654
While existing backdoor attacks have successfully infected multimodal contrastiv
e learning models such as CLIP they can be easily countered by specialized backd
oor defenses for MCL models. This paper reveals the threats in this practical sc
enario and introduces the BadCLIP attack which is resistant to backdoor detectio
n and model fine-tuning defenses. To achieve this we draw motivations from the p
erspective of the Bayesian rule and propose a dual-embedding guided framework fo
r backdoor attacks. Specifically we ensure that visual trigger patterns approxim
ate the textual target semantics in the embedding space making it challenging to
 detect the subtle parameter variations induced by backdoor learning on such nat
ural trigger patterns. Additionally we optimize the visual trigger patterns to a
lign the poisoned samples with target vision features in order to hinder backdoo
r unlearning through clean fine-tuning. Our experiments show a significant impro
vement in attack success rate (+45.3 % ASR) over current leading methods even ag
ainst state-of-the-art backdoor defenses highlighting our attack's effectiveness
 in various scenarios including downstream tasks. Our codes can be found at http
s://github.com/LiangSiyuan21/BadCLIP.
****
Dynamic Inertial Poser (DynaIP): Part-Based Motion Dynamics Learning for Enhance
d Human Pose Estimation with Sparse Inertial Sensors
Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, Ling Pei; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 1889-1899
This paper introduces a novel human pose estimation approach using sparse inerti
al sensors addressing the shortcomings of previous methods reliant on synthetic
data. It leverages a diverse array of real inertial motion capture data from dif
ferent skeleton formats to improve motion diversity and model generalization. Th
is method features two innovative components: a pseudo-velocity regression model
 for dynamic motion capture with inertial sensors and a part-based model dividin
g the body and sensor data into three regions each focusing on their unique char
acteristics. The approach demonstrates superior performance over state-of-the-ar
t models across five public datasets notably reducing pose error by 19% on the D
IP-IMU dataset thus representing a significant improvement in inertial sensor-ba
sed human pose estimation. Our codes are available at https://github.com/dx118/d
ynaip
****
Matching 2D Images in 3D: Metric Relative Pose from Metric Correspondences
Axel Barroso-Laguna, Sowmya Munukutla, Victor Adrian Prisacariu, Eric Brachmann;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2024, pp. 4852-4863
Given two images we can estimate the relative camera pose between them by establ
ishing image-to-image correspondences. Usually correspondences are 2D-to-2D and
the pose we estimate is defined only up to scale. Some applications aiming at in
stant augmented reality anywhere require scale-metric pose estimates and hence t
hey rely on external depth estimators to recover the scale. We present MicKey a
keypoint matching pipeline that is able to predict metric correspondences in 3D
camera space. By learning to match 3D coordinates across images we are able to i

nfer the metric relative pose without depth measurements. Depth measurements are also not required for training nor are scene reconstructions or image overlap information. MicKey is supervised only by pairs of images and their relative poses. MicKey achieves state-of-the-art performance on the Map-Free Relocalisation benchmark while requiring less supervision than competing approaches.
****
Efficient Vision-Language Pre-training by Cluster Masking
Zihao Wei, Zixuan Pan, Andrew Owens; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26815-26825
We propose a simple strategy for masking image patches during visual-language contrastive learning that improves the quality of the learned representations and the training speed. During each iteration of training we randomly mask clusters of visually similar image patches as measured by their raw pixel intensities. This provides an extra learning signal beyond the contrastive training itself since it forces a model to predict words for masked visual structures solely from context. It also speeds up training by reducing the amount of data used in each image. We evaluate the effectiveness of our model by pre-training on a number of benchmarks finding that it outperforms other masking strategies such as FLIP on the quality of the learned representation.
****
GraCo: Granularity-Controllable Interactive Segmentation
Yian Zhao, Kehan Li, Zesen Cheng, Pengchong Qiao, Xiawu Zheng, Rongrong Ji, Chang Liu, Li Yuan, Jie Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3501-3510
Interactive Segmentation (IS) segments specific objects or parts in the image according to user input. Current IS pipelines fall into two categories: single-granularity output and multi-granularity output. The latter aims to alleviate the spatial ambiguity present in the former. However the multi-granularity output pipeline suffers from limited interaction flexibility and produces redundant results. In this work we introduce Granularity-Controllable Interactive Segmentation (GraCo) a novel approach that allows precise control of prediction granularity by introducing additional parameters to input. This enhances the customization of the interactive system and eliminates redundancy while resolving ambiguity. Nevertheless the exorbitant cost of annotating multi-granularity masks and the lack of available datasets with granularity annotations make it difficult for models to acquire the necessary guidance to control output granularity. To address this problem we design an any-granularity mask generator that exploits the semantic property of the pre-trained IS model to automatically generate abundant mask-granularity pairs without requiring additional manual annotation. Based on these pairs we propose a granularity-controllable learning strategy that efficiently imparts the granularity controllability to the IS model. Extensive experiments on intricate scenarios at object and part levels demonstrate that our GraCo has significant advantages over previous methods. This highlights the potential of GraCo to be a flexible annotation tool capable of adapting to diverse segmentation scenarios. The project page: https://zhao-yian.github.io/GraCo.
****
M3-UDA: A New Benchmark for Unsupervised Domain Adaptive Fetal Cardiac Structure Detection
Bin Pu, Liwen Wang, Jiewen Yang, Guannan He, Xingbo Dong, Shengli Li, Ying Tan, Ming Chen, Zhe Jin, Kenli Li, Xiaomeng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11621-11630
The anatomical structure detection of fetal cardiac views is crucial for diagnosing fetal congenital heart disease. In practice there is a large domain gap between different hospitals' data such as the variable data quality due to differences in acquisition equipment. In addition accurate annotation information provided by obstetrician experts is always very costly or even unavailable. This study explores the unsupervised domain adaptive fetal cardiac structure detection issue. Existing unsupervised domain adaptive object detection (UDAOD) approaches mainly focus on detecting objects in natural scenes such as Foggy Cityscapes where the structural relationships of natural scenes are uncertain. Unlike all previou

s UDAOD scenarios we first collected a Fetal Cardiac Structure dataset from two hospital centers called FCS and proposed a multi-matching UDA approach (M3-UDA) including Histogram Matching (HM) Sub-structure Matching (SM) and Global-structure Matching (GM) to better transfer the topological knowledge of anatomical structure for UDA detection in medical scenarios. HM mitigates the domain gap between the source and target caused by pixel transformation. SM fuses the different angle information of the sub-structure to obtain the local topological knowledge for bridging the domain gap of the internal sub-structure. GM is designed to align the global topological knowledge of the whole organ from the source and target domain. Extensive experiments on our collected FCS and CardiacUDA and experimental results show that M3-UDA outperforms existing UDAOD studies significantly. All datasets and source code are available at : https://github.com/xmed-lab/M3-UDA
****
GPS-Gaussian: Generalizable Pixel-wise 3D Gaussian Splatting for Real-time Human Novel View Synthesis
Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19680-19690
We present a new approach termed GPS-Gaussian for synthesizing novel views of a character in a real-time manner. The proposed method enables 2K-resolution rendering under a sparse-view camera setting. Unlike the original Gaussian Splatting or neural implicit rendering methods that necessitate per-subject optimizations we introduce Gaussian parameter maps defined on the source views and regress directly Gaussian Splatting properties for instant novel view synthesis without any fine-tuning or optimization. To this end we train our Gaussian parameter regression module on a large amount of human scan data jointly with a depth estimation module to lift 2D parameter maps to 3D space. The proposed framework is fully differentiable and experiments on several datasets demonstrate that our method outperforms state-of-the-art methods while achieving an exceeding rendering speed.
****
Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding
Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, Li Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13700-13710
Large language models have demonstrated impressive universal capabilities across a wide range of open-ended tasks and have extended their utility to encompass multimodal conversations. However existing methods encounter challenges in effectively handling both image and video understanding particularly with limited visual tokens. In this work we introduce Chat-UniVi a Unified Vision-language model capable of comprehending and engaging in conversations involving images and videos through a unified visual representation. Specifically we employ a set of dynamic visual tokens to uniformly represent images and videos. This representation framework empowers the model to efficiently utilize a limited number of visual tokens to simultaneously capture the spatial details necessary for images and the comprehensive temporal relationship required for videos. Moreover we leverage a multi-scale representation enabling the model to perceive both high-level semantic concepts and low-level visual details. Notably Chat-UniVi is trained on a mixed dataset containing both images and videos allowing direct application to tasks involving both mediums without requiring any modifications. Extensive experimental results demonstrate that Chat-UniVi consistently outperforms even existing methods exclusively designed for either images or videos. Code is available at https://github.com/PKU-YuanGroup/Chat-UniVi.
****
MAGICK: A Large-scale Captioned Dataset from Matting Generated Images using Chroma Keying
Ryan D. Burgert, Brian L. Price, Jason Kuen, Yijun Li, Michael S. Ryoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22595-22604

We introduce MAGICK a large-scale dataset of generated objects with high-quality alpha mattes. While image generation methods have produced segmentations they cannot generate alpha mattes with accurate details in hair fur and transparencies. This is likely due to the small size of current alpha matting datasets and the difficulty in obtaining ground-truth alpha. We propose a scalable method for synthesizing images of objects with high-quality alpha that can be used as a ground-truth dataset. A key idea is to generate objects on a single-colored background so chroma keying approaches can be used to extract the alpha. However this faces several challenges including that current text-to-image generation methods cannot create images that can be easily chroma keyed and that chroma keying is an underconstrained problem that generally requires manual intervention for high-quality results. We address this using a combination of generation and alpha extraction methods. Using our method we generate a dataset of 150000 objects with alpha. We show the utility of our dataset by training an alpha-to-rgb generation method that outperforms baselines. Please see our project website at https://ryanndagreat.github.io/MAGICK/.
****
Video Super-Resolution Transformer with Masked Inter&Intra-Frame Attention
Xingyu Zhou, Leheng Zhang, Xiaorui Zhao, Keze Wang, Leida Li, Shuhang Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25399-25408
Recently Vision Transformer has achieved great success in recovering missing details in low-resolution sequences i.e. the video super-resolution (VSR) task. Despite its superiority in VSR accuracy the heavy computational burden as well as the large memory footprint hinder the deployment of Transformer-based VSR models on constrained devices. In this paper we address the above issue by proposing a novel feature-level masked processing framework: VSR with Masked Intra and inter-frame Attention (MIA-VSR). The core of MIA-VSR is leveraging feature-level temporal continuity between adjacent frames to reduce redundant computations and make more rational use of previously enhanced SR features. Concretely we propose an intra-frame and inter-frame attention block which takes the respective roles of past features and input features into consideration and only exploits previously enhanced features to provide supplementary information. In addition an adaptive block-wise mask prediction module is developed to skip unimportant computations according to feature similarity between adjacent frames. We conduct detailed ablation studies to validate our contributions and compare the proposed method with recent state-of-the-art VSR approaches. The experimental results demonstrate that MIA-VSR improves the memory and computation efficiency over state-of-the-art methods without trading off PSNR accuracy. The code is available at https://github.com/LabShuHangGU/MIA-VSR.
****
Token Transformation Matters: Towards Faithful Post-hoc Explanation for Vision Transformer
Junyi Wu, Bin Duan, Weitai Kang, Hao Tang, Yan Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10926-10935
While Transformers have rapidly gained popularity in various computer vision applications post-hoc explanations of their internal mechanisms remain largely unexplored. Vision Transformers extract visual information by representing image regions as transformed tokens and integrating them via attention weights. However existing post-hoc explanation methods merely consider these attention weights neglecting crucial information from the transformed tokens which fails to accurately illustrate the rationales behind the models' predictions. To incorporate the influence of token transformation into interpretation we propose TokenTM a novel post-hoc explanation method that utilizes our introduced measurement of token transformation effects. Specifically we quantify token transformation effects by measuring changes in token lengths and correlations in their directions pre- and post-transformation. Moreover we develop initialization and aggregation rules to integrate both attention weights and token transformation effects across all layers capturing holistic token contributions throughout the model. Experimental r

esults on segmentation and perturbation tests demonstrate the superiority of our proposed TokenTM compared to state-of-the-art Vision Transformer explanation me thods.
****
Bayesian Differentiable Physics for Cloth Digitalization
Deshan Gong, Ningtao Mao, He Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11841-11851
We propose a new method for cloth digitalization. Deviating from existing methods which learn from data captured under relatively casual settings we propose to learn from data captured in strictly tested measuring protocols and find plausible physical parameters of the cloths. However such data is currently absent so we first propose a new dataset with accurate cloth measurements. Further the data size is considerably smaller than the ones in current deep learning due to the nature of the data capture process. To learn from small data we propose a new Bayesian differentiable cloth model to estimate the complex material heterogeneity of real cloths. It can provide highly accurate digitalization from very limited data samples. Through exhaustive evaluation and comparison we show our method is accurate in cloth digitalization efficient in learning from limited data samples and general in capturing material variations. Code and data are available in: https://github.com/realcrane/Bayesian-Differentiable-Physics-for-Cloth-Digitalization
****
G-HOP: Generative Hand-Object Prior for Interaction Reconstruction and Grasp Synthesis
Yufei Ye, Abhinav Gupta, Kris Kitani, Shubham Tulsiani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1911-1920
We propose G-HOP a denoising diffusion based generative prior for hand-object interactions that allows modeling both the 3D object and a human hand conditioned on the object category. To learn a 3D spatial diffusion model that can capture this joint distribution we represent the human hand via a skeletal distance field to obtain a representation aligned with the (latent) signed distance field for the object. We show that this hand-object prior can then serve as a generic guidance to facilitate other tasks like reconstruction from interaction clip and human grasp synthesis. We believe that our model trained by aggregating several diverse real-world interaction datasets spanning 155 categories represents a first approach that allows jointly generating both hand and object. Our empirical evaluations demonstrate the benefit of this joint prior in video-based reconstruction and human grasp synthesis outperforming current task-specific baselines.
****
Higher-order Relational Reasoning for Pedestrian Trajectory Prediction
Sungjune Kim, Hyung-gun Chi, Hyerin Lim, Karthik Ramani, Jinkyu Kim, Sangpil Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15251-15260
Social relations have substantial impacts on the potential trajectories of each individual. Modeling these dynamics has been a central solution for more precise and accurate trajectory forecasting. However previous works ignore the importance of `social depth' meaning the influences flowing from different degrees of social relations. In this work we propose HighGraph a graph-based pedestrian relational reasoning method that captures the higher-order dynamics of social interactions. First we construct a collision-aware relation graph based on the agents' observed trajectories. Upon this graph structure we build our core module that aggregates the agent features from diverse social distances. As a result the network is able to model complex social relations thereby yielding more accurate and socially acceptable trajectories. Our HighGraph is a plug-and-play module that can be easily applied to any current trajectory predictors. Extensive experiments with ETH/UCY and SDD datasets demonstrate that our HighGraph noticeably improves the previous state-of-the-art baselines both quantitatively and qualitatively.
****

SurroundSDF: Implicit 3D Scene Understanding Based on Signed Distance Field
Lizhe Liu, Bohua Wang, Hongwei Xie, Daqi Liu, Li Liu, Zhiqiang Tian, Kuiyuan Yang, Bing Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21614-21623
Vision-centric 3D environment understanding is both vital and challenging for autonomous driving systems. Recently object-free methods have attracted considerable attention. Such methods perceive the world by predicting the semantics of discrete voxel grids but fail to construct continuous and accurate obstacle surfaces. To this end in this paper we propose SurroundSDF to implicitly predict the signed distance field (SDF) and semantic field for the continuous perception from surround images. Specifically we introduce a query-based approach and utilize SDF constrained by the Eikonal formulation to accurately describe the surfaces of obstacles. Furthermore considering the absence of precise SDF ground truth we propose a novel weakly supervised paradigm for SDF referred to as the Sandwich Eikonal formulation which emphasizes applying correct and dense constraints on both sides of the surface thereby enhancing the perceptual accuracy of the surface. Experiments suggest that our method achieves SOTA for both occupancy prediction and 3D scene reconstruction tasks on the nuScenes dataset.
****
Contrastive Denoising Score for Text-guided Latent Diffusion Image Editing
Hyelin Nam, Gihyun Kwon, Geon Yeong Park, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9192-9201
With the remarkable advent of text-to-image diffusion models image editing methods have become more diverse and continue to evolve. A promising recent approach in this realm is Delta Denoising Score (DDS) - an image editing technique based on Score Distillation Sampling (SDS) framework that leverages the rich generative prior of text-to-image diffusion models. However relying solely on the difference between scoring functions is insufficient for preserving specific structural elements from the original image a crucial aspect of image editing. To address this here we present an embarrassingly simple yet very powerful modification of DDS called Contrastive Denoising Score (CDS) for latent diffusion models (LDM). Inspired by the similarities and differences between DDS and the contrastive learning for unpaired image-to-image translation(CUT) we introduce a straightforward approach using CUT loss within the DDS framework. Rather than employing auxiliary networks as in the original CUT approach we leverage the intermediate features of LDM specifically those from the self-attention layers which possesses rich spatial information. Our approach enables zero-shot image-to-image translation and neural radiance field (NeRF) editing achieving structural correspondence between the input and output while maintaining content controllability. Qualitative results and comparisons demonstrates the effectiveness of our proposed method.
****
Neural Point Cloud Diffusion for Disentangled 3D Shape and Appearance Generation
Philipp Schröppel, Christopher Wewer, Jan Eric Lenssen, Eddy Ilg, Thomas Brox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8785-8794
Controllable generation of 3D assets is important for many practical applications like content creation in movies games and engineering as well as in AR/VR. Recently diffusion models have shown remarkable results in generation quality of 3D objects. However none of the existing models enable disentangled generation to control the shape and appearance separately. For the first time we present a suitable representation for 3D diffusion models to enable such disentanglement by introducing a hybrid point cloud and neural radiance field approach. We model a diffusion process over point positions jointly with a high-dimensional feature space for a local density and radiance decoder. While the point positions represent the coarse shape of the object the point features allow modeling the geometry and appearance details. This disentanglement enables us to sample both independently and therefore to control both separately. Our approach sets a new state of the art in generation compared to previous disentanglement-capable methods by reduced FID scores of 30-90% and is on-par with other non-disentanglement-capable

state-of-the art methods.
****
RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection
Ximiao Zhang, Min Xu, Xiuzhuang Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16699-16708
Self-supervised feature reconstruction methods have shown promising advances in industrial image anomaly detection and localization. Despite this progress these methods still face challenges in synthesizing realistic and diverse anomaly samples as well as addressing the feature redundancy and pre-training bias of pre-trained feature. In this work we introduce RealNet a feature reconstruction network with realistic synthetic anomaly and adaptive feature selection. It is incorporated with three key innovations: First we propose Strength-controllable Diffusion Anomaly Synthesis (SDAS) a diffusion process-based synthesis strategy capable of generating samples with varying anomaly strengths that mimic the distribution of real anomalous samples. Second we develop Anomaly-aware Features Selection (AFS) a method for selecting representative and discriminative pre-trained feature subsets to improve anomaly detection performance while controlling computational costs. Third we introduce Reconstruction Residuals Selection (RRS) a strategy that adaptively selects discriminative residuals for comprehensive identification of anomalous regions across multiple levels of granularity. We assess RealNet on four benchmark datasets and our results demonstrate significant improvements in both Image AUROC and Pixel AUROC compared to the current state-of-the-art methods. The code data and models are available at https://github.com/cnulab/RealNet.
****
Outdoor Scene Extrapolation with Hierarchical Generative Cellular Automata
Dongsu Zhang, Francis Williams, Zan Gojcic, Karsten Kreis, Sanja Fidler, Young Min Kim, Amlan Kar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20145-20154
We aim to generate fine-grained 3D geometry from large-scale sparse LiDAR scans abundantly captured by autonomous vehicles (AV). Contrary to prior work on AV scene completion we aim to extrapolate fine geometry from unlabeled and beyond spatial limits of LiDAR scans taking a step towards generating realistic high-resolution simulation-ready 3D street environments. We propose hierarchical Generative Cellular Automata (hGCA) a spatially scalable conditional 3D generative model which grows geometry recursively with local kernels following GCAs in a coarse-to-fine manner equipped with a light-weight planner to induce global consistency. Experiments on synthetic scenes show that hGCA generates plausible scene geometry with higher fidelity and completeness compared to state-of-the-art baselines. Our model generalizes strongly from sim-to-real qualitatively outperforming baselines on the Waymo-open dataset. We also show anecdotal evidence of the ability to create novel objects from real-world geometric cues even when trained on limited synthetic content.
****
Instruct 4D-to-4D: Editing 4D Scenes as Pseudo-3D Scenes Using 2D Diffusion
Linzhan Mou, Jun-Kun Chen, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20176-20185
This paper proposes Instruct 4D-to-4D that achieves 4D awareness and spatial-temporal consistency for 2D diffusion models to generate high-quality instruction-guided dynamic scene editing results. Traditional applications of 2D diffusion models in dynamic scene editing often result in inconsistency primarily due to their inherent frame-by-frame editing methodology. Addressing the complexities of extending instruction-guided editing to 4D our key insight is to treat a 4D scene as a pseudo-3D scene decoupled into two sub-problems: achieving temporal consistency in video editing and applying these edits to the pseudo-3D scene. Following this we first enhance the Instruct-Pix2Pix (IP2P) model with an anchor-aware attention module for batch processing and consistent editing. Additionally we integrate optical flow-guided appearance propagation in a sliding window fashion for more precise frame-to-frame editing and incorporate depth-based projection to

manage the extensive data of pseudo-3D scenes followed by iterative editing to a
chieve convergence. We extensively evaluate our approach in various scenes and e
diting instructions and demonstrate that it achieves spatially and temporally co
nsistent editing results with significantly enhanced detail and sharpness over t
he prior art. Notably Instruct 4D-to-4D is general and applicable to both monocu
lar and challenging multi-camera scenes.
****
VAREN: Very Accurate and Realistic Equine Network
Silvia Zuffi, Ylva Mellbin, Ci Li, Markus Hoeschle, Hedvig Kjellström, Senya Pol
ikovsky, Elin Hernlund, Michael J. Black; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5374-5383
Data-driven three-dimensional parametric shape models of the human body have gai
ned enormous popularity both for the analysis of visual data and for the generat
ion of synthetic humans. Following a similar approach for animals does not scale
 to the multitude of existing animal species not to mention the difficulty of ac
cessing subjects to scan in 3D. However we argue that for domestic species of gr
eat importance like the horse it is a highly valuable investment to put effort i
nto gathering a large dataset of real 3D scans and learn a realistic 3D articula
ted shape model. We introduce VAREN a novel 3D articulated parametric shape mode
l learned from 3D scans of many real horses. VAREN bridges synthesis and analysi
s tasks as the generated model instances have unprecedented realism while being
able to represent horses of different sizes and shapes. Differently from previou
s body models VAREN has two resolutions an anatomical skeleton and interpretable
 learned pose-dependent deformations which are related to the body muscles. We s
how with experiments that this formulation has superior performance with respect
 to previous strategies for modeling pose-dependent deformations in the human bo
dy case while also being more compact and allowing an analysis of the relationsh
ip between articulation and muscle deformation during articulated motion.
****
Photo-SLAM: Real-time Simultaneous Localization and Photorealistic Mapping for M
onocular Stereo and RGB-D Cameras
Huajian Huang, Longwei Li, Hui Cheng, Sai-Kit Yeung; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21584-2
1593
The integration of neural rendering and the SLAM system recently showed promisin
g results in joint localization and photorealistic view reconstruction. However
existing methods fully relying on implicit representations are so resource-hungr
y that they cannot run on portable devices which deviates from the original inte
ntion of SLAM. In this paper we present Photo-SLAM a novel SLAM framework with a
 hyper primitives map. Specifically we simultaneously exploit explicit geometric
 features for localization and learn implicit photometric features to represent
the texture information of the observed environment. In addition to actively den
sifying hyper primitives based on geometric features we further introduce a Gaus
sian-Pyramid-based training method to progressively learn multi-level features e
nhancing photorealistic mapping performance. The extensive experiments with mono
cular stereo and RGB-D datasets prove that our proposed system Photo-SLAM signif
icantly outperforms current state-of-the-art SLAM systems for online photorealis
tic mapping e.g. PSNR is 30% higher and rendering speed is hundreds of times fas
ter in the Replica dataset. Moreover the Photo-SLAM can run at real-time speed u
sing an embedded platform such as Jetson AGX Orin showing the potential of robot
ics applications. Project Page and code: https://huajianup.github.io/research/Ph
oto-SLAM/.
****
SD-DiT: Unleashing the Power of Self-supervised Discrimination in Diffusion Tran
sformer
Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, Chang Wen Chen
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2024, pp. 8435-8445
Diffusion Transformer (DiT) has emerged as the new trend of generative diffusion
 models on image generation. In view of extremely slow convergence in typical Di

T recent breakthroughs have been driven by mask strategy that significantly improves the training efficiency of DiT with additional intra-image contextual learning. Despite this progress mask strategy still suffers from two inherent limitations: (a) training-inference discrepancy and (b) fuzzy relations between mask reconstruction & generative diffusion process resulting in sub-optimal training of DiT. In this work we address these limitations by novelly unleashing the self-supervised discrimination knowledge to boost DiT training. Technically we frame our DiT in a teacher-student manner. The teacher-student discriminative pairs are built on the diffusion noises along the same Probability Flow Ordinary Differential Equation (PF-ODE). Instead of applying mask reconstruction loss over both DiT encoder and decoder we decouple DiT encoder and decoder to separately tackle discriminative and generative objectives. In particular by encoding discriminative pairs with student and teacher DiT encoders a new discriminative loss is designed to encourage the inter-image alignment in the self-supervised embedding space. After that student samples are fed into student DiT decoder to perform the typical generative diffusion task. Extensive experiments are conducted on ImageNet dataset and our method achieves a competitive balance between training cost and generative capacity.
****
Multi-modal Instruction Tuned LLMs with Fine-grained Visual Perception
Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, Xuansong Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13980-13990
Multimodal Large Language Model (MLLMs) leverages Large Language Models as a cognitive framework for diverse visual-language tasks. Recent efforts have been made to equip MLLMs with visual perceiving and grounding capabilities. However there still remains a gap in providing fine-grained pixel-level perceptions and extending interactions beyond text-specific inputs. In this work we propose \bf AnyRef a general MLLM model that can generate pixel-wise object perceptions and natural language descriptions from multi-modality references such as texts boxes images or audio. This innovation empowers users with greater flexibility to engage with the model beyond textual and regional prompts without modality-specific designs. Through our proposed refocusing mechanism the generated grounding output is guided to better focus on the referenced object implicitly incorporating additional pixel-level supervision. This simple modification utilizes attention scores generated during the inference of LLM eliminating the need for extra computations while exhibiting performance enhancements in both grounding masks and referring expressions. With only publicly available training data our model achieves state-of-the-art results across multiple benchmarks including diverse modality referring segmentation and region-level referring expression generation. Code and models are available at https://github.com/jwh97nn/AnyRef
****
ProMotion: Prototypes As Motion Learners
Yawen Lu, Dongfang Liu, Qifan Wang, Cheng Han, Yiming Cui, Zhiwen Cao, Xueling Zhang, Yingjie Victor Chen, Heng Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28109-28119
In this work we introduce ProMotion a unified prototypical transformer-based framework engineered to model fundamental motion tasks. ProMotion offers a range of compelling attributes that set it apart from current task-specific paradigms. 1. We adopt a prototypical perspective establishing a unified paradigm that harmonizes disparate motion learning approaches. This novel paradigm streamlines the architectural design enabling the simultaneous assimilation of diverse motion information. 2. We capitalize on a dual mechanism involving the feature denoiser and the prototypical learner to decipher the intricacies of motion. This approach effectively circumvents the pitfalls of ambiguity in pixel-wise feature matching significantly bolstering the robustness of motion representation. We demonstrate a profound degree of transferability across distinct motion patterns. This inherent versatility reverberates robustly across a comprehensive spectrum of both 2D and 3D downstream tasks. Empirical results demonstrate that outperforms various well-known specialized architectures achieving 0.54 and 0.054 Abs Rel error

on the Sintel and KITTI depth datasets 1.04 and 2.01 average endpoint error on the clean and final pass of Sintel flow benchmark and 4.30 F1-all error on the KITTI flow benchmark. For its efficacy we hope our work can catalyze a paradigm shift in universal models in computer vision.
****
SpatialTracker: Tracking Any 2D Pixels in 3D Space
Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20406-20417
Recovering dense and long-range pixel motion in videos is a challenging problem. Part of the difficulty arises from the 3D-to-2D projection process leading to occlusions and discontinuities in the 2D motion domain. While 2D motion can be intricate we posit that the underlying 3D motion can often be simple and low-dimensional. In this work we propose to estimate point trajectories in 3D space to mitigate the issues caused by image projection. Our method named SpatialTracker lifts 2D pixels to 3D using monocular depth estimators represents the 3D content of each frame efficiently using a triplane representation and performs iterative updates using a transformer to estimate 3D trajectories. Tracking in 3D allows us to leverage as-rigid-as possible(ARAP) constraints while simultaneously learning a rigidity embedding that clusters pixels into different rigid parts. Extensive evaluation shows that our approach achieves state-of-the-art tracking performance both qualitatively and quantitatively particularly in chal- lenging scenarios such as out-of-plane rotation. And our project page is available at https://henry123-boy.github.io/SpaTracker/.
****
LaMPilot: An Open Benchmark Dataset for Autonomous Driving with Language Model Programs
Yunsheng Ma, Can Cui, Xu Cao, Wenqian Ye, Peiran Liu, Juanwu Lu, Amr Abdelraouf, Rohit Gupta, Kyungtae Han, Aniket Bera, James M. Rehg, Ziran Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15141-15151
Autonomous driving (AD) has made significant strides in recent years. However existing frameworks struggle to interpret and execute spontaneous user instructions such as "overtake the car ahead." Large Language Models (LLMs) have demonstrated impressive reasoning capabilities showing potential to bridge this gap. In this paper we present LaMPilot a novel framework that integrates LLMs into AD systems enabling them to follow user instructions by generating code that leverages established functional primitives. We also introduce LaMPilot-Bench the first benchmark dataset specifically designed to quantitatively evaluate the efficacy of language model programs in AD. Adopting the LaMPilot framework we conduct extensive experiments to assess the performance of off-the-shelf LLMs on LaMPilot-Bench. Our results demonstrate the potential of LLMs in handling diverse driving scenarios and following user instructions in driving. To facilitate further research in this area we release our code and data at GitHub.com/PurdueDigitalTwin/LaMPilot.
****
MedBN: Robust Test-Time Adaptation against Malicious Test Samples
Hyejin Park, Jeongyeon Hwang, Sunung Mun, Sangdon Park, Jungseul Ok; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5997-6007
Test-time adaptation (TTA) has emerged as a promising solution to address performance decay due to unforeseen distribution shifts between training and test data. While recent TTA methods excel in adapting to test data variations such adaptability exposes a model to vulnerability against malicious examples an aspect that has received limited attention. Previous studies have uncovered security vulnerabilities within TTA even when a small proportion of the test batch is maliciously manipulated. In response to the emerging threat we propose median batch normalization (MedBN) leveraging the robustness of the median for statistics estimation within the batch normalization layer during test-time inference. Our method is algorithm-agnostic thus allowing seamless integration with existing TTA frame

works. Our experimental results on benchmark datasets including CIFAR10-C CIFAR100-C and ImageNet-C consistently demonstrate that MedBN outperforms existing approaches in maintaining robust performance across different attack scenarios encompassing both instant and cumulative attacks. Through extensive experiments we show that our approach sustains the performance even in the absence of attacks achieving a practical balance between robustness and performance.
****
Unsupervised Gaze Representation Learning from Multi-view Face Images
Yiwei Bao, Feng Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1419-1428
Annotating gaze is an expensive and time-consuming endeavor requiring costly eye-trackers or complex geometric calibration procedures. Although some eye-based unsupervised gaze representation learning methods have been proposed the quality of gaze representation extracted by these methods degrades severely when the head pose is large. In this paper we present the Multi-View Dual-Encoder (MV-DE) a framework designed to learn gaze representations from unlabeled multi-view face images. Through the proposed Dual-Encoder architecture and the multi-view gaze representation swapping strategy the MV-DE successfully disentangles gaze from general facial information and derives gaze representations closely tied to the subject's eyeball rotation without gaze label. Experimental results illustrate that the gaze representations learned by the MV-DE can be used in downstream tasks including gaze estimation and redirection. Gaze estimation results indicates that the proposed MV-DE displays notably higher robustness to uncontrolled head movements when compared to state-of-the-art (SOTA) unsupervised learning methods.
****
FairDeDup: Detecting and Mitigating Vision-Language Fairness Disparities in Semantic Dataset Deduplication
Eric Slyman, Stefan Lee, Scott Cohen, Kushal Kafle; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13905-13916
Recent dataset deduplication techniques have demonstrated that content-aware dataset pruning can dramatically reduce the cost of training Vision-Language Pretrained (VLP) models without significant performance losses compared to training on the original dataset. These results have been based on pruning commonly used image-caption datasets collected from the web -- datasets that are known to harbor harmful social biases that may then be codified in trained models. In this work we evaluate how deduplication affects the prevalence of these biases in the resulting trained models and introduce an easy-to-implement modification to the recent SemDeDup algorithm that can reduce the negative effects that we observe. When examining CLIP-style models trained on deduplicated variants of LAION-400M we find our proposed FairDeDup algorithm consistently leads to improved fairness metrics over SemDeDup on the FairFace and FACET datasets while maintaining zero-shot performance on CLIP benchmarks.
****
CrossMAE: Cross-Modality Masked Autoencoders for Region-Aware Audio-Visual Pre-Training
Yuxin Guo, Siyang Sun, Shuailei Ma, Kecheng Zheng, Xiaoyi Bao, Shijie Ma, Wei Zou, Yun Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26721-26731
Learning joint and coordinated features across modalities is essential for many audio-visual tasks. Existing pre-training methods primarily focus on global information neglecting fine-grained features and positions leading to suboptimal performance in dense prediction tasks. To address this issue we take a further step towards region-aware audio-visual pre-training and propose CrossMAE which excels in Cross-modality interaction and region alignment. Specifically we devise two masked autoencoding (MAE) pretext tasks at both pixel and embedding levels namely Cross-Conditioned Reconstruction and Cross-Embedding Reconstruction. Taking the visual modality as an example (the same goes for audio) in Cross-Conditioned Reconstruction the visual modality reconstructs the input image pixels conditioned on audio Attentive Tokens. As for the more challenging Cross-Embedding Recons

truction unmasked visual tokens reconstruct complete audio features under the guidance of learnable queries implying positional information which effectively enhances the interaction between modalities and exploits fine-grained semantics. Experimental results demonstrate that CrossMAE achieves state-of-the-art performance not only in classification and retrieval but also in dense prediction tasks. Furthermore we dive into the mechanism of modal interaction and region alignment of CrossMAE highlighting the effectiveness of the proposed components.
****
Osprey: Pixel Understanding with Visual Instruction Tuning
Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, Jianke Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28202-28211
Multimodal large language models (MLLMs) have recently achieved impressive general-purpose vision-language capabilities through visual instruction tuning. However current MLLMs primarily focus on image-level or box-level understanding falling short in achieving fine-grained vision-language alignment at pixel level. Besides the lack of mask-based instruction data limits their advancements. In this paper we propose Osprey a mask-text instruction tuning approach to extend MLLMs by incorporating fine-grained mask regions into language instruction aiming at achieving pixel-wise visual understanding. To achieve this goal we first meticulously curate a mask-based region-text dataset with 724K samples and then design a vision-language model by injecting pixel-level representation into LLM. Specifically Osprey adopts a convolutional CLIP backbone as the vision encoder and employs a mask-aware visual extractor to extract precise visual mask features from high resolution input. Experimental results demonstrate Osprey's superiority in various region understanding tasks showcasing its new capability for pixel-level instruction tuning. In particular Osprey can be integrated with Segment Anything Model (SAM) seamlessly to obtain multi-granularity semantics. The source code dataset and demo can be found at https://github.com/CircleRadon/Osprey.
****
Modality-agnostic Domain Generalizable Medical Image Segmentation by Multi-Frequency in Multi-Scale Attention
Ju-Hyeon Nam, Nur Suriza Syazwany, Su Jung Kim, Sang-Chul Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11480-11491
Generalizability in deep neural networks plays a pivotal role in medical image segmentation. However deep learning-based medical image analyses tend to overlook the importance of frequency variance which is critical element for achieving a model that is both modality-agnostic and domain-generalizable. Additionally various models fail to account for the potential information loss that can arise from multi-task learning under deep supervision a factor that can impair the model's representation ability. To address these challenges we propose a Modality-agnostic Domain Generalizable Network (MADGNet) for medical image segmentation which comprises two key components: a Multi-Frequency in Multi-Scale Attention (MFMSA) block and Ensemble Sub-Decoding Module (E-SDM). The MFMSA block refines the process of spatial feature extraction particularly in capturing boundary features by incorporating multi-frequency and multi-scale features thereby offering informative cues for tissue outline and anatomical structures. Moreover we propose E-SDM to mitigate information loss in multi-task learning with deep supervision especially during substantial upsampling from low resolution. We evaluate the segmentation performance of MADGNet across six modalities and fifteen datasets. Through extensive experiments we demonstrate that MADGNet consistently outperforms state-of-the-art models across various modalities showcasing superior segmentation performance. This affirms MADGNet as a robust solution for medical image segmentation that excels in diverse imaging scenarios. Our MADGNet code is available in GitHub Link.
****
Few-shot Learner Parameterization by Diffusion Time-steps
Zhongqi Yue, Pan Zhou, Richang Hong, Hanwang Zhang, Qianru Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,

pp. 23263-23272
Even when using large multi-modal foundation models few-shot learning is still challenging -- if there is no proper inductive bias it is nearly impossible to keep the nuanced class attributes while removing the visually prominent attributes that spuriously correlate with class labels. To this end we find an inductive bias that the time-steps of a Diffusion Model (DM) can isolate the nuanced class attributes i.e. as the forward diffusion adds noise to an image at each time-step nuanced attributes are usually lost at an earlier time-step than the spurious attributes that are visually prominent. Building on this we propose Time-step Few-shot (TiF) learner. We train class-specific low-rank adapters for a text-conditioned DM to make up for the lost attributes such that images can be accurately reconstructed from their noisy ones given a prompt. Hence at a small time-step the adapter and prompt are essentially a parameterization of only the nuanced class attributes. For a test image we can use the parameterization to only extract the nuanced class attributes for classification. TiF learner significantly outperforms OpenCLIP and its adapters on a variety of fine-grained and customized few-shot learning tasks. Codes are in https://github.com/yue-zhongqi/tif.
****
Auto MC-Reward: Automated Dense Reward Design with Large Language Models for Minecraft
Hao Li, Xue Yang, Zhaokai Wang, Xizhou Zhu, Jie Zhou, Yu Qiao, Xiaogang Wang, Hongsheng Li, Lewei Lu, Jifeng Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16426-16435
Many reinforcement learning environments (e.g. Minecraft) provide only sparse rewards that indicate task completion or failure with binary values. The challenge in exploration efficiency in such environments makes it difficult for reinforcement-learning-based agents to learn complex tasks. To address this this paper introduces an advanced learning system named Auto MC-Reward that leverages Large Language Models (LLMs) to automatically design dense reward functions thereby enhancing the learning efficiency. Auto MC-Reward consists of three important components: Reward Designer Reward Critic and Trajectory Analyzer. Given the environment information and task descriptions the Reward Designer first design the reward function by coding an executable Python function with predefined observation inputs. Then our Reward Critic will be responsible for verifying the code checking whether the code is self-consistent and free of syntax and semantic errors. Further the Trajectory Analyzer summarizes possible failure causes and provides refinement suggestions according to collected trajectories. In the next round Reward Designer will further refine and iterate the dense reward function based on feedback. Experiments demonstrate a significant improvement in the success rate and learning efficiency of our agents in complex tasks in Minecraft such as obtaining diamond with the efficient ability to avoid lava and efficiently explore trees and animals that are sparse in the plains biome.
****
GenFlow: Generalizable Recurrent Flow for 6D Pose Refinement of Novel Objects
Sungphill Moon, Hyeontae Son, Dongcheol Hur, Sangwook Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10039-10049
Despite the progress of learning-based methods for 6D object pose estimation the trade-off between accuracy and scalability for novel objects still exists. Specifically previous methods for novel objects do not make good use of the target object's 3D shape information since they focus on generalization by processing the shape indirectly making them less effective. We present GenFlow an approach that enables both accuracy and generalization to novel objects with the guidance of the target object's shape. Our method predicts optical flow between the rendered image and the observed image and refines the 6D pose iteratively. It boosts the performance by a constraint of the 3D shape and the generalizable geometric knowledge learned from an end-to-end differentiable system. We further improve our model by designing a cascade network architecture to exploit the multi-scale correlations and coarse-to-fine refinement. GenFlow ranked first on the unseen object pose estimation benchmarks in both the RGB and RGB-D cases. It also achieve

s performance competitive with existing state-of-the-art methods for the seen ob
ject pose estimation without any fine-tuning.
****
OrCo: Towards Better Generalization via Orthogonality and Contrast for Few-Shot Class-Incremental Learning
Noor Ahmed, Anna Kukleva, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28762-28771
Few-Shot Class-Incremental Learning (FSCIL) introduces a paradigm in which the p
roblem space expands with limited data. FSCIL methods inherently face the challe
nge of catastrophic forgetting as data arrives incrementally making models susce
ptible to overwriting previously acquired knowledge. Moreover given the scarcity
 of labeled samples available at any given time models may be prone to overfitti
ng and find it challenging to strike a balance between extensive pretraining and
 the limited incremental data. To address these challenges we propose the OrCo f
ramework built on two core principles: features' orthogonality in the representa
tion space and contrastive learning. In particular we improve the generalization
 of the embedding space by employing a combination of supervised and self-superv
ised contrastive losses during the pretraining phase. Additionally we introduce
OrCo loss to address challenges arising from data limitations during incremental
 sessions. Through feature space perturbations and orthogonality between classes
 the OrCo loss maximizes margins and reserves space for the following incrementa
l data. This in turn ensures the accommodation of incoming classes in the featur
e space without compromising previously acquired knowledge. Our experimental res
ults showcase state-of-the-art performance across three benchmark datasets inclu
ding mini-ImageNet CIFAR100 and CUB datasets. Code is available at https://githu
b.com/noorahmedds/OrCo
****
MuGE: Multiple Granularity Edge Detection
Caixia Zhou, Yaping Huang, Mengyang Pu, Qingji Guan, Ruoxi Deng, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25952-25962
Edge segmentation is well-known to be subjective due to personalized annotation
styles and preferred granularity. However most existing deterministic edge detec
tion methods produce only a single edge map for one input image. We argue that g
enerating multiple edge maps is more reasonable than generating a single one con
sidering the subjectivity and ambiguity of the edges. Thus motivated in this pap
er we propose multiple granularity edge detection called MuGE which can produce
a wide range of edge maps from approximate object contours to fine texture edges
. Specifically we first propose to design an edge granularity network to estimat
e the edge granularity from an individual edge annotation. Subsequently to guide
 the generation of diversified edge maps we integrate such edge granularity into
 the multi-scale feature maps in the spatial domain. Meanwhile we decompose the
feature maps into low-frequency and high-frequency parts where the encoded edge
granularity is further fused into the high-frequency part to achieve more precis
e control over the details of the produced edge maps. Compared to previous metho
ds MuGE is able to not only generate multiple edge maps at different controllabl
e granularities but also achieve a competitive performance on the BSDS500 and Mu
lticue benchmark datasets.
****
Real-World Efficient Blind Motion Deblurring via Blur Pixel Discretization
Insoo Kim, Jae Seok Choi, Geonseok Seo, Kinam Kwon, Jinwoo Shin, Hyong-Euk Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25879-25888
As recent advances in mobile camera technology have enabled the capability to ca
pture high-resolution images such as 4K images the demand for an efficient deblu
rring model handling large motion has increased. In this paper we discover that
the image residual errors i.e. blur-sharp pixel differences can be grouped into
some categories according to their motion blur type and how complex their neighb
oring pixels are. Inspired by this we decompose the deblurring (regression) task
 into blur pixel discretization (pixel-level blur classification) and discrete-t

o-continuous conversion (regression with blur class map) tasks. Specifically we generate the discretized image residual errors by identifying the blur pixels and then transform them to a continuous form which is computationally more efficient than naively solving the original regression problem with continuous values. Here we found that the discretization result i.e. blur segmentation map remarkably exhibits visual similarity with the image residual errors. As a result our efficient model shows comparable performance to state-of-the-art methods in realistic benchmarks while our method is up to 10 times computationally more efficient.

****

EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, Wen-Huang Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26596-26605

Visual Instruction Tuning represents a novel learning paradigm involving the fine-tuning of pre-trained language models using task-specific instructions. This paradigm shows promising zero-shot results in various natural language processing tasks but is still unexplored in vision emotion understanding. In this work we focus on enhancing the model's proficiency in understanding and adhering to instructions related to emotional contexts. Initially we identify key visual clues critical to visual emotion recognition. Subsequently we introduce a novel GPT-assisted pipeline for generating emotion visual instruction data effectively addressing the scarcity of annotated instruction data in this domain. Expanding on the groundwork established by InstructBLIP our proposed EmoVIT architecture incorporates emotion-specific instruction data leveraging the powerful capabilities of Large Language Models to enhance performance. Through extensive experiments our model showcases its proficiency in emotion classification adeptness in affective reasoning and competence in comprehending humor. The comparative analysis provides a robust benchmark for Emotion Visual Instruction Tuning in the era of LLMs providing valuable insights and opening avenues for future exploration in this domain. Our code is available at https://github.com/aimmemotion/EmoVIT.

****

Learning to Count without Annotations

Lukas Knobel, Tengda Han, Yuki M. Asano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22924-22934

While recent supervised methods for reference-based object counting continue to improve the performance on benchmark datasets they have to rely on small datasets due to the cost associated with manually annotating dozens of objects in images. We propose UnCounTR a model that can learn this task without requiring any manual annotations. To this end we construct "Self-Collages" images with various pasted objects as training samples that provide a rich learning signal covering arbitrary object types and counts. Our method builds on existing unsupervised representations and segmentation techniques to successfully demonstrate for the first time the ability of reference-based counting without manual supervision. Our experiments show that our method not only outperforms simple baselines and generic models such as FasterRCNN and DETR but also matches the performance of supervised counting models in some domains.

****

Logarithmic Lenses: Exploring Log RGB Data for Image Classification

Bruce A. Maxwell, Sumegha Singhania, Avnish Patel, Rahul Kumar, Heather Fryling, Sihan Li, Haonan Sun, Ping He, Zewen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17470-17479

The design of deep network architectures and training methods in computer vision has been well-explored. However in almost all cases the images have been used as provided with little exploration of pre-processing steps beyond normalization and data augmentation. Virtually all images posted on the web or captured by devices are processed for viewing by humans. Is the pipeline used for humans also best for use by computers and deep networks? The human visual system uses logarithmic sensors; differences and sums correspond to ratios and products. Features in log space will be invariant to intensity changes and robust to color balance c

hanges. Log RGB space also reveals structure that is corrupted by typical pre-processing. We explore using linear and log RGB data for training standard backbone architectures on an image classification task using data derived directly from RAW images to guarantee its integrity. We found that networks trained on log RGB data exhibit improved performance on an unmodified test set and invariance to intensity and color balance modifications without additional training or data augmentation. Furthermore we found that the gains from using high quality log data could also be partially or fully realized from data in 8-bit sRGB-JPG format by inverting the sRGB transform and taking the log. These results imply existing databases may benefit from this type of pre-processing. While working with log data we found it was critical to retain the integrity of the log relationships and that networks using log data train best with meta-parameters different than those used for sRGB or linear data. Finally we introduce a new 10-category 10k RAW image data set (RAW10) for image classification and other purposes to enable further the exploration of log RGB as an input format for deep networks in computer vision.
****
AEROBLADE: Training-Free Detection of Latent Diffusion Images Using Autoencoder Reconstruction Error
Jonas Ricker, Denis Lukovnikov, Asja Fischer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9130-9140
With recent text-to-image models anyone can generate deceptively realistic images with arbitrary contents fueling the growing threat of visual disinformation. A key enabler for generating high-resolution images with low computational cost has been the development of latent diffusion models (LDMs). In contrast to conventional diffusion models LDMs perform the denoising process in the low-dimensional latent space of a pre-trained autoencoder (AE) instead of the high-dimensional image space. Despite their relevance the forensic analysis of LDMs is still in its infancy. In this work we propose AEROBLADE a novel detection method which exploits an inherent component of LDMs: the AE used to transform images between image and latent space. We find that generated images can be more accurately reconstructed by the AE than real images allowing for a simple detection approach based on the reconstruction error. Most importantly our method is easy to implement and does not require any training yet nearly matches the performance of detectors that rely on extensive training. We empirically demonstrate that AEROBLADE is effective against state-of-the-art LDMs including Stable Diffusion and Midjourney. Beyond detection our approach allows for the qualitative analysis of images which can be leveraged for identifying inpainted regions. We release our code and data at https://github.com/jonasricker/aeroblade.
****
Scaled Decoupled Distillation
Shicai Wei, Chunbo Luo, Yang Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15975-15983
Logit knowledge distillation attracts increasing attention due to its practicality in recent studies. However it often suffers inferior performance compared to the feature knowledge distillation. In this paper we argue that existing logit-based methods may be sub-optimal since they only leverage the global logit output that couples multiple semantic knowledge. This may transfer ambiguous knowledge to the student and mislead its learning. To this end we propose a simple but effective method i.e. Scale Decoupled Distillation (SDD) for logit knowledge distillation. SDD decouples the global logit output into multiple local logit outputs and establishes distillation pipelines for them. This helps the student to mine and inherit fine-grained and unambiguous logit knowledge. Moreover the decoupled knowledge can be further divided into consistent and complementary logit knowledge that transfers the semantic information and sample ambiguity respectively. By increasing the weight of complementary parts SDD can guide the student to focus more on ambiguous samples improving its discrimination ability. Extensive experiments on several benchmark datasets demonstrate the effectiveness of SDD for wide teacher-student pairs especially in the fine-grained classification task. Code is available at: \href https://github.com/shicaiwei123/SDD-CVPR2024 https:/

/github.com/shicaiwei123/SDD-CVPR2024
****
NARUTO: Neural Active Reconstruction from Uncertain Target Observations
Ziyue Feng, Huangying Zhan, Zheng Chen, Qingan Yan, Xiangyu Xu, Changjiang Cai, Bing Li, Qilun Zhu, Yi Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21572-21583
We present NARUTO a neural active reconstruction system that combines a hybrid neural representation with uncertainty learning enabling high-fidelity surface reconstruction. Our approach leverages a multi-resolution hash-grid as the mapping backbone chosen for its exceptional convergence speed and capacity to capture high-frequency local features. The centerpiece of our work is the incorporation of an uncertainty learning module that dynamically quantifies reconstruction uncertainty while actively reconstructing the environment. By harnessing learned uncertainty we propose a novel uncertainty aggregation strategy for goal searching and efficient path planning. Our system autonomously explores by targeting uncertain observations and reconstructs environments with remarkable completeness and fidelity. We also demonstrate the utility of this uncertainty-aware approach by enhancing SOTA neural SLAM systems through an active ray sampling strategy. Extensive evaluations of NARUTO in various environments using an indoor scene simulator confirm its superior performance and state-of-the-art status in active reconstruction as evidenced by its impressive results on benchmark datasets like Replica and MP3D.
****
Point2CAD: Reverse Engineering CAD Models from 3D Point Clouds
Yujia Liu, Anton Obukhov, Jan Dirk Wegner, Konrad Schindler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3763-3772
Computer-Aided Design (CAD) model reconstruction from point clouds is an important problem at the intersection of computer vision graphics and machine learning; it saves the designer significant time when iterating on in-the-wild objects. Recent advancements in this direction achieve relatively reliable semantic segmentation but still struggle to produce an adequate topology of the CAD model. In this work we analyze the current state of the art for that ill-posed task and identify shortcomings of existing methods. We propose a hybrid analytic-neural reconstruction scheme that bridges the gap between segmented point clouds and structured CAD models and can be readily combined with different segmentation backbones. Moreover to power the surface fitting stage we propose a novel implicit neural representation of freeform surfaces driving up the performance of our overall CAD reconstruction scheme. We extensively evaluate our method on the popular ABC benchmark of CAD models and set a new state-of-the-art for that dataset. Code is available at https://github.com/YujiaLiu76/point2cad.
****
Learnable Earth Parser: Discovering 3D Prototypes in Aerial Scans
Romain Loiseau, Elliot Vincent, Mathieu Aubry, Loic Landrieu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27874-27884
We propose an unsupervised method for parsing large 3D scans of real-world scenes with easily-interpretable shapes. This work aims to provide a practical tool for analyzing 3D scenes in the context of aerial surveying and mapping without the need for user annotations. Our approach is based on a probabilistic reconstruction model that decomposes an input 3D point cloud into a small set of learned prototypical 3D shapes. The resulting reconstruction is visually interpretable and can be used to perform unsupervised instance and low-shot semantic segmentation of complex scenes. We demonstrate the usefulness of our model on a novel dataset of seven large aerial LiDAR scans from diverse real-world scenarios. Our approach outperforms state-of-the-art unsupervised methods in terms of decomposition accuracy while remaining visually interpretable. Our code and dataset are available at https://romainloiseau.fr/learnable-earth-parser/.
****
NeRFiller: Completing Scenes via Generative 3D Inpainting

Ethan Weber, Aleksander Holynski, Varun Jampani, Saurabh Saxena, Noah Snavely, A
bhishek Kar, Angjoo Kanazawa; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2024, pp. 20731-20741
We propose NeRFiller an approach that completes missing portions of a 3D capture
 via generative 3D inpainting using off-the-shelf 2D visual generative models. O
ften parts of a captured 3D scene or object are missing due to mesh reconstructi
on failures or a lack of observations (e.g. contact regions such as the bottom o
f objects or hard-to-reach areas). We approach this challenging 3D inpainting pr
oblem by leveraging a 2D inpainting diffusion model. We identify a surprising be
havior of these models where they generate more 3D consistent inpaints when imag
es form a 2x2 grid and show how to generalize this behavior to more than four im
ages. We then present an iterative framework to distill these inpainted regions
into a single consistent 3D scene. In contrast to related works we focus on comp
leting scenes rather than deleting foreground objects and our approach does not
require tight 2D object masks or text. We compare our approach to relevant basel
ines adapted to our setting on a variety of scenes where NeRFiller creates the m
ost 3D consistent and plausible scene completions. Our project page is at https:
//ethanweber.me/nerfiller/.
****
Cloud-Device Collaborative Learning for Multimodal Large Language Models
Guanqun Wang, Jiaming Liu, Chenxuan Li, Yuan Zhang, Junpeng Ma, Xinyu Wei, Kevin
 Zhang, Maurice Chong, Renrui Zhang, Yijiang Liu, Shanghang Zhang; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 12646-12655
The burgeoning field of Multimodal Large Language Models (MLLMs) has exhibited r
emarkable performance in diverse tasks such as captioning commonsense reasoning
and visual scene understanding. However the deployment of these large-scale MLLM
s on client devices is hindered by their extensive model parameters leading to a
 notable decline in generalization capabilities when these models are compressed
 for device deployment. Addressing this challenge we introduce a Cloud-Device Co
llaborative Continual Adaptation framework designed to enhance the performance o
f compressed device-deployed MLLMs by leveraging the robust capabilities of clou
d-based larger-scale MLLMs. Our framework is structured into three key component
s: a device-to-cloud uplink for efficient data transmission cloud-based knowledg
e adaptation and an optimized cloud-to-device downlink for model deployment. In
the uplink phase we employ an Uncertainty-guided Token Sampling (UTS) strategy t
o effectively filter out-of-distribution tokens thereby reducing transmission co
sts and improving training efficiency. On the cloud side we propose Adapter-base
d Knowledge Distillation (AKD) method to transfer refined knowledge from large-s
cale to compressed pocket-size MLLMs. Furthermore we propose a Dynamic Weight up
date Compression (DWC) strategy for the downlink which adaptively selects and qu
antizes updated weight parameters enhancing transmission efficiency and reducing
 the representational disparity between cloud and device models. Extensive exper
iments on several multimodal benchmarks demonstrate the superiority of our propo
sed framework over prior Knowledge Distillation and device-cloud collaboration m
ethods. Notably we also validate the feasibility of our approach to real-world e
xperiments.
****
KD-DETR: Knowledge Distillation for Detection Transformer with Consistent Distil
lation Points Sampling
Yu Wang, Xin Li, Shengzhao Weng, Gang Zhang, Haixiao Yue, Haocheng Feng, Junyu H
an, Errui Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2024, pp. 16016-16025
DETR is a novel end-to-end transformer architecture object detector which signif
icantly outperforms classic detectors when scaling up. In this paper we focus on
 the compression of DETR with knowledge distillation. While knowledge distillati
on has been well-studied in classic detectors there is a lack of researches on h
ow to make it work effectively on DETR. We first provide experimental and theore
tical analysis to point out that the main challenge in DETR distillation is the
lack of consistent distillation points. Distillation points refer to the corresp

onding inputs of the predictions for student to mimic which have different formu
lations in CNN detector and DETR and reliable distillation requires sufficient d
istillation points which are consistent between teacher and student. Based on th
is observation we propose the first general knowledge distillation paradigm for
DETR(KD-DETR) with consistent distillation points sampling for both homogeneous
and heterogeneous distillation. Specifically we decouple detection and distillat
ion tasks by introducing a set of specialized object queries to construct distil
lation points for DETR. We further propose a general-to-specific distillation po
ints sampling strategy to explore the extensibility of KD-DETR. Extensive experi
ments validate the effectiveness and generalization of KD-DETR. For both single-
scale DAB-DETR and multis-scale Deformable DETR and DINO KD-DETR boost the perfo
rmance of student model with improvements of 2.6%-5.2%. We further extend KD-DET
R to heterogeneous distillation and achieves 2.1% improvement by distilling the
knowledge from DINO to Faster R-CNN with ResNet-50 which is comparable with homo
geneous distillation methods.
****
Absolute Pose from One or Two Scaled and Oriented Features
Jonathan Ventura, Zuzana Kukelova, Torsten Sattler, Dániel Baráth; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 20870-20880
Keypoints used for image matching often include an estimate of the feature scale
 and orientation. While recent work has demonstrated the advantages of using fea
ture scales and orientations for relative pose estimation relatively little work
 has considered their use for absolute pose estimation. We introduce minimal sol
utions for absolute pose from two oriented feature correspondences in the genera
l case or one scaled and oriented correspondence given a known vertical directio
n. Nowadays assuming a known direction is not particularly restrictive as modern
 consumer devices such as smartphones or drones are equipped with Inertial Measu
rement Units (IMU) that provide the gravity direction by default. Compared to tr
aditional absolute pose methods requiring three point correspondences our solver
s need a smaller minimal sample reducing the cost and complexity of robust estim
ation. Evaluations on large-scale and public real datasets demonstrate the advan
tage of our methods for fast and accurate localization in challenging conditions
. Code is available at https://github.com/danini/absolute-pose-from-oriented-and
-scaled-features .
****
Source-Free Domain Adaptation with Frozen Multimodal Foundation Model
Song Tang, Wenxin Su, Mao Ye, Xiatian Zhu; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23711-23720
Source-Free Domain Adaptation (SFDA) aims to adapt a source model for a target d
omain with only access to unlabeled target training data and the source model pr
etrained on a supervised source domain. Relying on pseudo labeling and/or auxili
ary supervision conventional methods are inevitably error-prone. To mitigate thi
s limitation in this work we for the first time explore the potentials of off-th
e-shelf vision-language (ViL) multimodal models (e.g. CLIP) with rich whilst het
erogeneous knowledge. We find that directly applying the ViL model to the target
 domain in a zero-shot fashion is unsatisfactory as it is not specialized for th
is particular task but largely generic. To make it task specific we propose a no
vel Distilling multImodal Foundation mOdel (DIFO) approach. Specifically DIFO al
ternates between two steps during adaptation: (i) Customizing the ViL model by m
aximizing the mutual information with the target model in a prompt learning mann
er (ii) Distilling the knowledge of this customized ViL model to the target mode
l. For more fine-grained and reliable distillation we further introduce two effe
ctive regularization terms namely most-likely category encouragement and predict
ive consistency. Extensive experiments show that DIFO significantly outperforms
the state-of-the-art alternatives. Code is here.
****
LocLLM: Exploiting Generalizable Human Keypoint Localization via Large Language
Model
Dongkai Wang, Shiyu Xuan, Shiliang Zhang; Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 614-623
The capacity of existing human keypoint localization models is limited by keypoint priors provided by the training data. To alleviate this restriction and pursue more general model this work studies keypoint localization from a different perspective by reasoning locations based on keypiont clues in text descriptions. We propose LocLLM the first Large-Language Model (LLM) based keypoint localization model that takes images and text instructions as inputs and outputs the desired keypoint coordinates. LocLLM leverages the strong reasoning capability of LLM and clues of keypoint type location and relationship in textual descriptions for keypoint localization. To effectively tune LocLLM we construct localization-based instruction conversations to connect keypoint description with corresponding coordinates in input image and fine-tune the whole model in a parameter-efficient training pipeline. LocLLM shows remarkable performance on standard 2D/3D keypoint localization benchmarks. Moreover incorporating language clues into the localization makes LocLLM show superior flexibility and generalizable capability in cross dataset keypoint localization and even detecting novel type of keypoints unseen during training.
****
MMA-Diffusion: MultiModal Attack on Diffusion Models
Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, Qiang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7737-7746
In recent years Text-to-Image (T2I) models have seen remarkable advancements gaining widespread adoption. However this progress has inadvertently opened avenues for potential misuse particularly in generating inappropriate or Not-Safe-For-Work (NSFW) content. Our work introduces MMA-Diffusion a framework that presents a significant and realistic threat to the security of T2I models by effectively circumventing current defensive measures in both open-source models and commercial online services. Unlike previous approaches MMA-Diffusion leverages both textual and visual modalities to bypass safeguards like prompt filters and post-hoc safety checkers thus exposing and highlighting the vulnerabilities in existing defense mechanisms. Our codes are available at https://github.com/cure-lab/MMA-Diffusion.
****
Benchmarking Audio Visual Segmentation for Long-Untrimmed Videos
Chen Liu, Peike Patrick Li, Qingtao Yu, Hongwei Sheng, Dadong Wang, Lincheng Li, Xin Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22712-22722
Existing audio-visual segmentation datasets typically focus on short-trimmed videos with only one pixel-map annotation for a per-second video clip. In contrast for untrimmed videos the sound duration start- and end-sounding time positions and visual deformation of audible objects vary significantly. Therefore we observed that current AVS models trained on trimmed videos might struggle to segment sounding objects in long videos. To investigate the feasibility of grounding audible objects in videos along both temporal and spatial dimensions we introduce the Long-Untrimmed Audio-Visual Segmentation dataset (LU-AVS) which includes precise frame-level annotations of sounding emission times and provides exhaustive mask annotations for all frames. Considering that pixel-level annotations are difficult to achieve in some complex scenes we also provide the bounding boxes to indicate the sounding regions. Specifically LU-AVS contains 10M mask annotations across 6.6K videos and 11M bounding box annotations across 7K videos. Compared with the existing datasets LU-AVS videos are on average 4 8 times longer with the silent duration being 3 15 times greater. Furthermore we try our best to adapt some baseline models that were originally designed for audio-visual-relevant tasks to examine the challenges of our newly curated LU-AVS. Through comprehensive evaluation we demonstrate the challenges of LU-AVS compared to the ones containing trimmed videos. Therefore LU-AVS provides an ideal yet challenging platform for evaluating audio-visual segmentation and localization on untrimmed long videos. The dataset is publicly available at: https://yenanliu.github.io/LU-AVS/.
****

EMCAD: Efficient Multi-scale Convolutional Attention Decoding for Medical Image Segmentation
Md Mostafijur Rahman, Mustafa Munir, Radu Marculescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11769-11779
An efficient and effective decoding mechanism is crucial in medical image segmentation especially in scenarios with limited computational resources. However these decoding mechanisms usually come with high computational costs. To address this concern we introduce EMCAD a new efficient multi-scale convolutional attention decoder designed to optimize both performance and computational efficiency. EMCAD leverages a unique multi-scale depth-wise convolution block significantly enhancing feature maps through multi-scale convolutions. EMCAD also employs channel spatial and grouped (large-kernel) gated attention mechanisms which are highly effective at capturing intricate spatial relationships while focusing on salient regions. By employing group and depth-wise convolution EMCAD is very efficient and scales well (e.g. only 1.91M parameters and 0.381G FLOPs are needed when using a standard encoder). Our rigorous evaluations across 12 datasets that belong to six medical image segmentation tasks reveal that EMCAD achieves state-of-the-art (SOTA) performance with 79.4% and 80.3% reduction in #Params and #FLOPs respectively. Moreover EMCAD's adaptability to different encoders and versatility across segmentation tasks further establish EMCAD as a promising tool advancing the field towards more efficient and accurate medical image analysis. Our implementation is available at https://github.com/SLDGroup/EMCAD.
****
VTQA: Visual Text Question Answering via Entity Alignment and Cross-Media Reasoning
Kang Chen, Xiangqian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27218-27227
Achieving the optimal form of Visual Question Answering mandates a profound grasp of understanding grounding and reasoning within the intersecting domains of vision and language. Traditional VQA benchmarks have predominantly focused on simplistic tasks such as counting visual attributes and object detection which do not necessitate intricate cross-modal information understanding and inference. Motivated by the need for a more comprehensive evaluation we introduce a novel dataset comprising 23781 questions derived from 10124 image-text pairs. Specifically the task of this dataset requires the model to align multimedia representations of the same entity to implement multi-hop reasoning between image and text and finally use natural language to answer the question. Furthermore we evaluate this VTQA dataset comparing the performance of both state-of-the-art VQA models and our proposed baseline model the Key Entity Cross-Media Reasoning Network (KECMRN). The VTQA task poses formidable challenges for traditional VQA models underscoring its intrinsic complexity. Conversely KECMRN exhibits a modest improvement signifying its potential in multimedia entity alignment and multi-step reasoning. Our analysis underscores the diversity difficulty and scale of the VTQA task compared to previous multimodal QA datasets. In conclusion we anticipate that this dataset will serve as a pivotal resource for advancing and evaluating models proficient in multimedia entity alignment multi-step reasoning and open-ended answer generation. Our dataset and code is available at https://visual-text-qa.github.io/.
****
QN-Mixer: A Quasi-Newton MLP-Mixer Model for Sparse-View CT Reconstruction
Ishak Ayad, Nicolas Larue, Mai K. Nguyen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25317-25326
Inverse problems span across diverse fields. In medical contexts computed tomography (CT) plays a crucial role in reconstructing a patient's internal structure presenting challenges due to artifacts caused by inherently ill-posed inverse problems. Previous research advanced image quality via post-processing and deep unrolling algorithms but faces challenges such as extended convergence times with ultra-sparse data. Despite enhancements resulting images often show significant artifacts limiting their effectiveness for real-world diagnostic applications. W

e aim to explore deep second-order unrolling algorithms for solving imaging inverse problems emphasizing their faster convergence and lower time complexity compared to common first-order methods like gradient descent. In this paper we introduce QN-Mixer an algorithm based on the quasi-Newton approach. We use learned parameters through the BFGS algorithm and introduce Incept-Mixer an efficient neural architecture that serves as a non-local regularization term capturing long-range dependencies within images. To address the computational demands typically associated with quasi-Newton algorithms that require full Hessian matrix computations we present a memory-efficient alternative. Our approach intelligently downsamples gradient information significantly reducing computational requirements while maintaining performance. The approach is validated through experiments on the sparse-view CT problem involving various datasets and scanning protocols and is compared with post-processing and deep unrolling state-of-the-art approaches. Our method outperforms existing approaches and achieves state-of-the-art performance in terms of SSIM and PSNR all while reducing the number of unrolling iterations required.
****
Learning CNN on ViT: A Hybrid Model to Explicitly Class-specific Boundaries for Domain Adaptation
Ba Hung Ngo, Nhat-Tuong Do-Tran, Tuan-Ngoc Nguyen, Hae-Gon Jeon, Tae Jong Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28545-28554
Most domain adaptation (DA) methods are based on either a convolutional neural networks (CNNs) or a vision transformers (ViTs). They align the distribution differences between domains as encoders without considering their unique characteristics. For instance ViT excels in accuracy due to its superior ability to capture global representations while CNN has an advantage in capturing local representations. This fact has led us to design a hybrid method to fully take advantage of both ViT and CNN called Explicitly Class-specific Boundaries (ECB). ECB learns CNN on ViT to combine their distinct strengths. In particular we leverage ViT's properties to explicitly find class-specific decision boundaries by maximizing the discrepancy between the outputs of the two classifiers to detect target samples far from the source support. In contrast the CNN encoder clusters target features based on the previously defined class-specific boundaries by minimizing the discrepancy between the probabilities of the two classifiers. Finally ViT and CNN mutually exchange knowledge to improve the quality of pseudo labels and reduce the knowledge discrepancies of these models. Compared to conventional DA methods our ECB achieves superior performance which verifies its effectiveness in this hybrid model. The project website can be found https://dotrannhattuong.github.io/ECB/website/.
****
A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions
Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, Adriana Romero-Soriano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26700-26709
Curation methods for massive vision-language datasets trade off between dataset size and quality. However even the highest quality of available curated captions are far too short to capture the rich visual detail in an image. To show the value of dense and highly-aligned image-text pairs we collect the Densely Captioned Images (DCI) dataset containing 8012 natural images human-annotated with mask-aligned descriptions averaging above 1000 words each. With precise and reliable captions associated with specific parts of an image we can evaluate vision-language models' (VLMs) understanding of image content with a novel task that matches each caption with its corresponding subcrop. As current models are often limited to 77 text tokens we also introduce a summarized version (sDCI) in which each caption length is limited. We show that modern techniques that make progress on standard benchmarks do not correspond with significant improvement on our sDCI based benchmark. Lastly we finetune CLIP using sDCI and show significant improvements over the baseline despite a small training set. By releasing the first huma

n annotated dense image captioning dataset we hope to enable the development of new benchmarks or fine-tuning recipes for the next generation of VLMs to come.
****
HanDiffuser: Text-to-Image Generation With Realistic Hand Appearances
Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2468-2479
Text-to-image generative models can generate high-quality humans but realism is lost when generating hands. Common artifacts include irregular hand poses shapes incorrect numbers of fingers and physically implausible finger orientations. To generate images with realistic hands we propose a novel diffusion-based architecture called HanDiffuser that achieves realism by injecting hand embeddings in the generative process. HanDiffuser consists of two components: a Text-to-Hand-Params diffusion model to generate SMPL-Body and MANO-Hand parameters from input text prompts and a Text-Guided Hand-Params-to-Image diffusion model to synthesize images by conditioning on the prompts and hand parameters generated by the previous component. We incorporate multiple aspects of hand representation including 3D shapes and joint-level finger positions orientations and articulations for robust learning and reliable performance during inference. We conduct extensive quantitative and qualitative experiments and perform user studies to demonstrate the efficacy of our method in generating images with high-quality hands.
****
Infinigen Indoors: Photorealistic Indoor Scenes using Procedural Generation
Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, Jia Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21783-21794
We introduce Infinigen Indoors a Blender-based procedural generator of photorealistic indoor scenes. It builds upon the existing Infinigen system which focuses on natural scenes but expands its coverage to indoor scenes by introducing a diverse library of procedural indoor assets including furniture architecture elements appliances and other day-to-day objects. It also introduces a constraint-based arrangement system which consists of a domain-specific language for expressing diverse constraints on scene composition and a solver that generates scene compositions that maximally satisfy the constraints. We provide an export tool that allows the generated 3D objects and scenes to be directly used for training embodied agents in real-time simulators such as Omniverse and Unreal. Infinigen Indoors is open-sourced under the BSD license. Please visit infinigen.org for code and videos.
****
MART: Masked Affective RepresenTation Learning via Masked Temporal Distribution Distillation
Zhicheng Zhang, Pancheng Zhao, Eunil Park, Jufeng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12830-12840
Limited training data is a long-standing problem for video emotion analysis (VEA). Existing works leverage the power of large-scale image datasets for transferring while failing to extract the temporal correlation of affective cues in the video. Inspired by psychology research and empirical theory we verify that the degree of emotion may vary in different segments of the video thus introducing the sentiment complementary and emotion intrinsic among temporal segments. We propose an MAE-style method for learning robust affective representation of videos via a masking termed MART. First we extract the affective cues of the lexicon and verify the extracted one by computing its matching score with video content in terms of sentiment and emotion scores alongside the temporal dimension. Then with the verified cues we propose masked affective modeling to recover temporal emotion distribution. We present temporal affective complementary learning that pulls the complementary part and pushes the intrinsic one of masked multimodal features where the constraint is set with cross-modal attention among features to mask the video and recover the degree of emotion among segments. Extensive experiment

s on five benchmarks show the superiority of our method in video sentiment analy sis video emotion recognition multimodal sentiment analysis and multimodal emoti on recognition.
****
MTLoRA: Low-Rank Adaptation Approach for Efficient Multi-Task Learning
Ahmed Agiza, Marina Neseem, Sherief Reda; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16196-16205
Adapting models pre-trained on large-scale datasets to a variety of downstream t asks is a common strategy in deep learning. Consequently parameter-efficient fin e-tuning methods have emerged as a promising way to adapt pre-trained models to different tasks while training only a minimal number of parameters. While most o f these methods are designed for single-task adaptation parameter-efficient trai ning in Multi-Task Learning (MTL) architectures is still unexplored. In this pap er we introduce MTLoRA a novel framework for parameter-efficient training of MTL models. MTLoRA employs Task-Agnostic and Task-Specific Low-Rank Adaptation modu les which effectively disentangle the parameter space in MTL fine-tuning thereby enabling the model to adeptly handle both task specialization and interaction w ithin MTL contexts. We applied MTLoRA to hierarchical-transformer-based MTL arch itectures adapting them to multiple downstream dense prediction tasks. Our exten sive experiments on the PASCAL dataset show that MTLoRA achieves higher accuracy on downstream tasks compared to fully fine-tuning the MTL model while reducing the number of trainable parameters by 3.6x. Furthermore MTLoRA establishes a Par eto-optimal trade-off between the number of trainable parameters and the accurac y of the downstream tasks outperforming current state-of-the-art parameter-effic ient training methods in both accuracy and efficiency.
****
Hierarchical Patch Diffusion Models for High-Resolution Video Generation
Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 7569-7579
Diffusion models have demonstrated remarkable performance in image and video syn thesis. However scaling them to high-resolution inputs is challenging and requir es restructuring the diffusion pipeline into multiple independent components lim iting scalability and complicating downstream applications. In this work we stud y patch diffusion models (PDMs) -- a diffusion paradigm which models the distrib ution of patches rather than whole inputs keeping up to 0.7% of the original pix els. This makes it very efficient during training and unlocks end-to-end optimiz ation on high-resolution videos. We improve PDMs in two principled ways. First t o enforce consistency between patches we develop deep context fusion -- an archi tectural technique that propagates the context information from low-scale to hig h-scale patches in a hierarchical manner. Second to accelerate training and infe rence we propose adaptive computation which allocates more network capacity and computation towards coarse image details. The resulting model sets a new state-o f-the-art FVD score of 66.32 and Inception Score of 87.68 in class-conditional v ideo generation on UCF-101 256x256 surpassing recent methods by more than 100%. Then we show that it can be rapidly fine-tuned from a base 36x64 low-resolution generator for high-resolution 64x288x512 text-to-video synthesis. To the best of our knowledge our model is the first diffusion-based architecture which is trai ned on such high resolutions entirely end-to-end. Project webpage: https://snap-research.github.io/hpdm.
****
Motion Blur Decomposition with Cross-shutter Guidance
Xiang Ji, Haiyang Jiang, Yinqiang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12534-12543
Motion blur is a frequently observed image artifact especially under insufficien t illumination where exposure time has to be prolonged so as to collect more pho tons for a bright enough image. Rather than simply removing such blurring effect s recent researches have aimed at decomposing a blurry image into multiple sharp images with spatial and temporal coherence. Since motion blur decomposition its elf is highly ambiguous priors from neighbouring frames or human annotation are

usually needed for motion disambiguation. In this paper inspired by the compleme ntary exposure characteristics of a global shutter (GS) camera and a rolling shu tter (RS) camera we propose to utilize the ordered scanline-wise delay in a roll ing shutter image to robustify motion decomposition of a single blurry image. To evaluate this novel dual imaging setting we construct a triaxial system to coll ect realistic data as well as a deep network architecture that explicitly addres ses temporal and contextual information through reciprocal branches for cross-sh utter motion blur decomposition. Experiment results have verified the effectiven ess of our proposed algorithm as well as the validity of our dual imaging settin g.
****

Scene-adaptive and Region-aware Multi-modal Prompt for Open Vocabulary Object De tection

Xiaowei Zhao, Xianglong Liu, Duorui Wang, Yajun Gao, Zhide Liu; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16741-16750

Open Vocabulary Object Detection (OVD) aims to detect objects from novel classes described by text inputs based on the generalization ability of trained classes . Existing methods mainly focus on transferring knowledge from large Vision and Language models (VLM) to detectors through knowledge distillation. However these approaches show weak ability in adapting to diverse classes and aligning betwee n the image-level pre-training and region-level detection thereby impeding effec tive knowledge transfer. Motivated by the prompt tuning we propose scene-adaptiv e and region-aware multi-modal prompts to address these issues by effectively ad apting class-aware knowledge from VLM to the detector at the region level. Speci fically to enhance the adaptability to diverse classes we design a scene-adaptiv e prompt generator from a scene perspective to consider both the commonality and diversity of the class distributions and formulate a novel selection mechanism to facilitate the acquisition of common knowledge across all classes and specifi c insights relevant to each scene. Meanwhile to bridge the gap between the pre-t rained model and the detector we present a region-aware multi-modal alignment mo dule which employs the region prompt to incorporate the positional information f or feature distillation and integrates textual prompts to align visual and lingu istic representations. Extensive experimental results demonstrate that the propo sed method significantly outperforms the state-of-the-art models on the OV-COCO and OV-LVIS datasets surpassing the current method by 3.0% mAP and 4.6% APr .
****

MimicDiffusion: Purifying Adversarial Perturbation via Mimicking Clean Diffusion Model

Kaiyu Song, Hanjiang Lai, Yan Pan, Jian Yin; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24665-24674

Deep neural networks (DNNs) are vulnerable to adversarial perturbation where an imperceptible perturbation is added to the image that can fool the DNNs. Diffusi on-based adversarial purification uses the diffusion model to generate a clean i mage against such adversarial attacks. Unfortunately the generative process of t he diffusion model is also inevitably affected by adversarial perturbation since the diffusion model is also a deep neural network where its input has adversari al perturbation. In this work we propose MimicDiffusion a new diffusion-based ad versarial purification technique that directly approximates the generative proce ss of the diffusion model with the clean image as input. Concretely we analyze t he differences between the guided terms using the clean image and the adversaria l sample. After that we first implement MimicDiffusion based on Manhattan distan ce. Then we propose two guidance to purify the adversarial perturbation and appr oximate the clean diffusion model. Extensive experiments on three image datasets including CIFAR-10 CIFAR-100 and ImageNet with three classifier backbones inclu ding WideResNet-70-16 WideResNet-28-10 and ResNet-50 demonstrate that MimicDiffu sion significantly performs better than the state-of-the-art baselines. On CIFAR -10 CIFAR-100 and ImageNet it achieves 92.67% 61.35% and 61.53% average robust a ccuracy which are 18.49% 13.23% and 17.64% higher respectively. The code is avai lable at https://github.com/psky1111/MimicDiffusion.

****
Neural Implicit Morphing of Face Images
Guilherme Schardong, Tiago Novello, Hallison Paz, Iurii Medvedev, Vinícius da Silva, Luiz Velho, Nuno Gonçalves; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7321-7330
Face morphing is a problem in computer graphics with numerous artistic and forensic applications. It is challenging due to variations in pose lighting gender and ethnicity. This task consists of a warping for feature alignment and a blending for a seamless transition between the warped images. We propose to leverage coord-based neural networks to represent such warpings and blendings of face images. During training we exploit the smoothness and flexibility of such networks by combining energy functionals employed in classical approaches without discretizations. Additionally our method is time-dependent allowing a continuous warping/blending of the images. During morphing inference we need both direct and inverse transformations of the time-dependent warping. The first (second) is responsible for warping the target (source) image into the source (target) image. Our neural warping stores those maps in a single network dismissing the need for inverting them. The results of our experiments indicate that our method is competitive with both classical and generative models under the lens of image quality and face-morphing detectors. Aesthetically the resulting images present a seamless blending of diverse faces not yet usual in the literature.
****
UniGS: Unified Representation for Image Generation and Segmentation
Lu Qi, Lehan Yang, Weidong Guo, Yu Xu, Bo Du, Varun Jampani, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6305-6315
This paper introduces a novel unified representation of diffusion models for image generation and segmentation. Specifically we use a colormap to represent entity-level masks addressing the challenge of varying entity numbers while aligning the representation closely with the image RGB domain. Two novel modules including the location-aware color palette and progressive dichotomy module are proposed to support our mask representation. On the one hand a location-aware palette guarantees the colors' consistency to entities' locations. On the other hand the progressive dichotomy module can efficiently decode the synthesized colormap to high-quality entity-level masks in a depth-first binary search without knowing the cluster numbers. To tackle the issue of lacking large-scale segmentation training data we employ an inpainting pipeline and then improve the flexibility of diffusion models across various tasks including inpainting image synthesis referring segmentation and entity segmentation. Comprehensive experiments validate the efficiency of our approach demonstrating comparable segmentation mask quality to state-of-the-art and adaptability to multiple tasks.
****
Robust Synthetic-to-Real Transfer for Stereo Matching
Jiawei Zhang, Jiahe Li, Lei Huang, Xiaohan Yu, Lin Gu, Jin Zheng, Xiao Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20247-20257
With advancements in domain generalized stereo matching networks models pre-trained on synthetic data demonstrate strong robustness to unseen domains. However few studies have investigated the robustness after fine-tuning them in real-world scenarios during which the domain generalization ability can be seriously degraded. In this paper we explore fine-tuning stereo matching networks without compromising their robustness to unseen domains. Our motivation stems from comparing Ground Truth (GT) versus Pseudo Label (PL) for fine-tuning: GT degrades but PL preserves the domain generalization ability. Empirically we find the difference between GT and PL implies valuable information that can regularize networks during fine-tuning. We also propose a framework to utilize this difference for fine-tuning consisting of a frozen Teacher an exponential moving average (EMA) Teacher and a Student network. The core idea is to utilize the EMA Teacher to measure what the Student has learned and dynamically improve GT and PL for fine-tuning. We integrate our framework with state-of-the-art networks and evaluate its effect

iveness on several real-world datasets. Extensive experiments show that our meth od effectively preserves the domain generalization ability during fine-tuning.
****
Instance-Aware Group Quantization for Vision Transformers
Jaehyeon Moon, Dohyung Kim, Junyong Cheon, Bumsub Ham; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16132 -16141
Post-training quantization (PTQ) is an efficient model compression technique tha t quantizes a pretrained full-precision model using only a small calibration set of unlabeled samples without retraining. PTQ methods for convolutional neural n etworks (CNNs) provide quantization results comparable to full-precision counter parts. Directly applying them to vision transformers (ViTs) however incurs sever e performance degradation mainly due to the differences in architectures between CNNs and ViTs. In particular the distribution of activations for each channel v ary drastically according to input instances making PTQ methods for CNNs inappro priate for ViTs. To address this we introduce instance-aware group quantization for ViTs (IGQ-ViT). To this end we propose to split the channels of activation m aps into multiple groups dynamically for each input instance such that activatio ns within each group share similar statistical properties. We also extend our sc heme to quantize softmax attentions across tokens. In addition the number of gro ups for each layer is adjusted to minimize the discrepancies between predictions from quantized and full-precision models under a bit-operation (BOP) constraint . We show extensive experimental results on image classification object detectio n and instance segmentation with various transformer architectures demonstrating the effectiveness of our approach.
****
A General and Efficient Training for Transformer via Token Expansion
Wenxuan Huang, Yunhang Shen, Jiao Xie, Baochang Zhang, Gaoqi He, Ke Li, Xing Sun , Shaohui Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2024, pp. 15783-15792
The remarkable performance of Vision Transformers (ViTs) typically requires an e xtremely large training cost. Existing methods have attempted to accelerate the training of ViTs yet typically disregard method universality with accuracy dropp ing. Meanwhile they break the training consistency of the original transformers including the consistency of hyper-parameters architecture and strategy which pr events them from being widely applied to different Transformer networks. In this paper we propose a novel token growth scheme Token Expansion (termed ToE) to ac hieve consistent training acceleration for ViTs. We introduce an "initialization -expansion-merging" pipeline to maintain the integrity of the intermediate featu re distribution of original transformers preventing the loss of crucial learnabl e information in the training process. ToE can not only be seamlessly integrated into the training and fine-tuning process of transformers (e.g. DeiT and LV-ViT) but also effective for efficient training frameworks (e.g. EfficientTrain) wit hout twisting the original training hyper-parameters architecture and introducin g additional training strategies. Extensive experiments demonstrate that ToE ach ieves about 1.3x faster for the training of ViTs in a lossless manner or even wi th performance gains over the full-token training baselines. Code is available a t https://github.com/Osilly/TokenExpansion.
****
GenZI: Zero-Shot 3D Human-Scene Interaction Generation
Lei Li, Angela Dai; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2024, pp. 20465-20474
Can we synthesize 3D humans interacting with scenes without learning from any 3D human-scene interaction data? We propose GenZI the first zero-shot approach to generating 3D human-scene interactions. Key to GenZI is our distillation of inte raction priors from large vision-language models (VLMs) which have learned a ric h semantic space of 2D human-scene compositions. Given a natural language descri ption and a coarse point location of the desired interaction in a 3D scene we fi rst leverage VLMs to imagine plausible 2D human interactions inpainted into mult iple rendered views of the scene. We then formulate a robust iterative optimizat

ion to synthesize the pose and shape of a 3D human model in the scene guided by consistency with the 2D interaction hypotheses. In contrast to existing learning-based approaches GenZI circumvents the conventional need for captured 3D interaction data and allows for flexible control of the 3D interaction synthesis with easy-to-use text prompts. Extensive experiments show that our zero-shot approach has high flexibility and generality making it applicable to diverse scene types including both indoor and outdoor environments.
****
Tyche: Stochastic In-Context Learning for Medical Image Segmentation
Marianne Rakic, Hallee E. Wong, Jose Javier Gonzalez Ortiz, Beth A. Cimini, John V. Guttag, Adrian V. Dalca; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11159-11173
Existing learning-based solutions to medical image segmentation have two important shortcomings. First for most new segmentation tasks a new model has to be trained or fine-tuned. This requires extensive resources and machine-learning expertise and is therefore often infeasible for medical researchers and clinicians. Second most existing segmentation methods produce a single deterministic segmentation mask for a given image. In practice however there is often considerable uncertainty about what constitutes the correct segmentation and different expert annotators will often segment the same image differently. We tackle both of these problems with Tyche a framework that uses a context set to generate stochastic predictions for previously unseen tasks without the need to retrain. Tyche differs from other in-context segmentation methods in two important ways. (1) We introduce a novel convolution block architecture that enables interactions among predictions. (2) We introduce in-context test-time augmentation a new mechanism to provide prediction stochasticity. When combined with appropriate model design and loss functions Tyche can predict a set of plausible diverse segmentation candidates for new or unseen medical images and segmentation tasks without the need to retrain. Code available at: https://tyche.csail.mit.edu/.
****
DiffAssemble: A Unified Graph-Diffusion Model for 2D and 3D Reassembly
Gianluca Scarpellini, Stefano Fiorini, Francesco Giuliari, Pietro Moreiro, Alessio Del Bue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28098-28108
Reassembly tasks play a fundamental role in many fields and multiple approaches exist to solve specific reassembly problems. In this context we posit that a general unified model can effectively address them all irrespective of the input data type (image 3D etc.). We introduce DiffAssemble a Graph Neural Network (GNN)-based architecture that learns to solve reassembly tasks using a diffusion model formulation. Our method treats the elements of a set whether pieces of 2D patch or 3D object fragments as nodes of a spatial graph. Training is performed by introducing noise into the position and rotation of the elements and iteratively denoising them to reconstruct the coherent initial pose. DiffAssemble achieves state-of-the-art (SOTA) results in most 2D and 3D reassembly tasks and is the first learning-based approach that solves 2D puzzles for both rotation and translation. Furthermore we highlight its remarkable reduction in run-time performing 11 times faster than the quickest optimization-based method for puzzle solving.
****
NeISF: Neural Incident Stokes Field for Geometry and Material Estimation
Chenhao Li, Taishi Ono, Takeshi Uemori, Hajime Mihara, Alexander Gatto, Hajime Nagahara, Yusuke Moriuchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21434-21445
Multi-view inverse rendering is the problem of estimating the scene parameters such as shapes materials or illuminations from a sequence of images captured under different viewpoints. Many approaches however assume single light bounce and thus fail to recover challenging scenarios like inter-reflections. On the other hand simply extending those methods to consider multi-bounced light requires more assumptions to alleviate the ambiguity. To address this problem we propose Neural Incident Stokes Fields (NeISF) a multi-view inverse rendering framework that reduces ambiguities using polarization cues. The primary motivation for using po

larization cues is that it is the accumulation of multi-bounced light providing rich information about geometry and material. Based on this knowledge the proposed incident Stokes field efficiently models the accumulated polarization effect with the aid of an original physically-based differentiable polarimetric renderer. Lastly experimental results show that our method outperforms the existing works in synthetic and real scenarios.
****
Training-Free Open-Vocabulary Segmentation with Offline Diffusion-Augmented Prototype Generation
Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3689-3698
Open-vocabulary semantic segmentation aims at segmenting arbitrary categories expressed in textual form. Previous works have trained over large amounts of image-caption pairs to enforce pixel-level multimodal alignments. However captions provide global information about the semantics of a given image but lack direct localization of individual concepts. Further training on large-scale datasets inevitably brings significant computational costs. In this paper we propose FreeDA a training-free diffusion-augmented method for open-vocabulary semantic segmentation which leverages the ability of diffusion models to visually localize generated concepts and local-global similarities to match class-agnostic regions with semantic classes. Our approach involves an offline stage in which textual-visual reference embeddings are collected starting from a large set of captions and leveraging visual and semantic contexts. At test time these are queried to support the visual matching process which is carried out by jointly considering class-agnostic regions and global semantic similarities. Extensive analyses demonstrate that FreeDA achieves state-of-the-art performance on five datasets surpassing previous methods by more than 7.0 average points in terms of mIoU and without requiring any training. Our source code is available at https://aimagelab.github.io/freeda/.
****
YOLO-World: Real-Time Open-Vocabulary Object Detection
Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16901-16911
The You Only Look Once (YOLO) series of detectors have established themselves as efficient and practical tools. However their reliance on predefined and trained object categories limits their applicability in open scenarios. Addressing this limitation we introduce YOLO-World an innovative approach that enhances YOLO with open-vocabulary detection capabilities through vision-language modeling and pre-training on large-scale datasets. Specifically we propose a new Re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN) and region-text contrastive loss to facilitate the interaction between visual and linguistic information. Our method excels in detecting a wide range of objects in a zero-shot manner with high efficiency. On the challenging LVIS dataset YOLO-World achieves 35.4 AP with 52.0 FPS on V100 which outperforms many state-of-the-art methods in terms of both accuracy and speed. Furthermore the fine-tuned YOLO-World achieves remarkable performance on several downstream tasks including object detection and open-vocabulary instance segmentation. Code and models are available at https://github.com/AILab-CVC/YOLO-World
****
ViT-Lens: Towards Omni-modal Representations
Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26647-26657
Aiming to advance AI agents large foundation models significantly improve reasoning and instruction execution yet the current focus on vision and language neglects the potential of perceiving diverse modalities in open-world environments. However the success of data-driven vision and language models is costly or even infeasible to be reproduced for rare modalities. In this paper we present ViT-Len

s that facilitates efficient omni-modal representation learning by perceiving novel modalities with a pretrained ViT and aligning them to a pre-defined space. Specifically the modality-specific lens is tuned to project any-modal signals to an intermediate embedding space which are then processed by a strong ViT with pre-trained visual knowledge. The encoded representations are optimized toward aligning with the modal-independent space pre-defined by off-the-shelf foundation models. ViT-Lens provides a unified solution for representation learning of increasing modalities with two appealing advantages: (i) Unlocking the great potential of pretrained ViTs to novel modalities effectively with efficient data regime; (ii) Enabling emergent downstream capabilities through modality alignment and shared ViT parameters. We tailor ViT-Lens to learn representations for 3D point cloud depth audio tactile and EEG and set new state-of-the-art results across various understanding tasks such as zero-shot classification. By seamlessly integrating ViT-Lens into Multimodal Foundation Models we enable Any-modality to Text and Image Generation in a zero-shot manner. Code and models are available at https://github.com/TencentARC/ViT-Lens.
****
Cross-Dimension Affinity Distillation for 3D EM Neuron Segmentation
Xiaoyu Liu, Miaomiao Cai, Yinda Chen, Yueyi Zhang, Te Shi, Ruobing Zhang, Xuejin Chen, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11104-11113
Accurate 3D neuron segmentation from electron microscopy (EM) volumes is crucial for neuroscience research. However the complex neuron morphology often leads to over-merge and over-segmentation results. Recent advancements utilize 3D CNNs to predict a 3D affinity map with improved accuracy but suffer from two challenges: high computational cost and limited input size especially for practical deployment for large-scale EM volumes. To address these challenges we propose a novel method to leverage lightweight 2D CNNs for efficient neuron segmentation. Our method employs a 2D Y-shape network to generate two embedding maps from adjacent 2D sections which are then converted into an affinity map by measuring their embedding distance. While the 2D network better captures pixel dependencies inside sections with larger input sizes it overlooks inter-section dependencies. To overcome this we introduce a cross-dimension affinity distillation (CAD) strategy that transfers inter-section dependency knowledge from a 3D teacher network to the 2D student network by ensuring consistency between their output affinity maps. Additionally we design a feature grafting interaction (FGI) module to enhance knowledge transfer by grafting embedding maps from the 2D student onto those from the 3D teacher. Extensive experiments on multiple EM neuron segmentation datasets including a newly built one by ourselves demonstrate that our method achieves superior performance over state-of-the-art methods with only 1/20 inference latency.
****
HUGS: Human Gaussian Splats
Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, Anurag Ranjan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 505-515
Recent advances in neural rendering have improved both training and rendering times by orders of magnitude. While these methods demonstrate state-of-the-art quality and speed they are designed for photogrammetry of static scenes and do not generalize well to freely moving humans in the environment. In this work we introduce Human Gaussian Splats (HUGS) that represents an animatable human together with the scene using 3D Gaussian Splatting (3DGS). Our method takes only a monocular video with a small number of (50-100) frames and it automatically learns to disentangle the static scene and a fully animatable human avatar within 30 minutes. We utilize the SMPL body model to initialize the human Gaussians. To capture details that are not modeled by SMPL (e.g cloth hairs) we allow the 3D Gaussians to deviate from the human body model. Utilizing 3D Gaussians for animated humans brings new challenges including the artifacts created when articulating the Gaussians. We propose to jointly optimize the linear blend skinning weights to coordinate the movements of individual Gaussians during animation. Our approach e

nables novel-pose synthesis of human and novel view synthesis of both the human and the scene. We achieve state-of-the-art rendering quality with a rendering speed of 60 FPS while being 100x faster to train over previous work.
****
GeoChat: Grounded Large Vision-Language Model for Remote Sensing
Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27831-27840
Recent advancements in Large Vision-Language Models (VLMs) have shown great promise in natural image domains allowing users to hold a dialogue about given visual content. However such general-domain VLMs perform poorly for Remote Sensing (RS) scenarios leading to inaccurate or fabricated information when presented with RS domain-specific queries. Such a behavior emerges due to the unique challenges introduced by RS imagery. For example to handle high-resolution RS imagery with diverse scale changes across categories and many small objects region-level reasoning is necessary alongside holistic scene interpretation. Furthermore the lack of domain-specific multimodal instruction following data as well as strong backbone models for RS make it hard for the models to align their behavior with user queries. To address these limitations we propose GeoChat - the first versatile remote sensing VLM that offers multitask conversational capabilities with high-resolution RS images. Specifically GeoChat can not only answer image-level queries but also accepts region inputs to hold region-specific dialogue. Furthermore it can visually ground objects in its responses by referring to their spatial coordinates. To address the lack of domain-specific datasets we generate a novel RS multimodal instruction-following dataset by extending image-text pairs from existing diverse RS datasets. Leveraging this rich dataset we fine-tune our remote sensing VLM based on the LLaVA-1.5 architecture. We establish a comprehensive benchmark for RS multitask conversations and compare with a number of baseline methods. GeoChat demonstrates robust zero-shot performance on various remote sensing tasks e.g. image and region captioning visual question answering scene classification visually grounded conversations and referring object detection. Our codes will be open-sourced.
****
PhysPT: Physics-aware Pretrained Transformer for Estimating Human Dynamics from Monocular Videos
Yufei Zhang, Jeffrey O. Kephart, Zijun Cui, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2305-2317
While current methods have shown promising progress on estimating 3D human motion from monocular videos their motion estimates are often physically unrealistic because they mainly consider kinematics. In this paper we introduce Physics-aware Pretrained Transformer (PhysPT) which improves kinematics-based motion estimates and infers motion forces. PhysPT exploits a Transformer encoder-decoder backbone to effectively learn human dynamics in a self-supervised manner. Moreover it incorporates physics principles governing human motion. Specifically we build a physics-based body representation and contact force model. We leverage them to impose novel physics-inspired training losses (i.e. force loss contact loss and Euler-Lagrange loss) enabling PhysPT to capture physical properties of the human body and the forces it experiences. Experiments demonstrate that once trained PhysPT can be directly applied to kinematics-based estimates to significantly enhance their physical plausibility and generate favourable motion forces. Furthermore we show that these physically meaningful quantities translate into improved accuracy of an important downstream task: human action recognition.
****
Producing and Leveraging Online Map Uncertainty in Trajectory Prediction
Xunjiang Gu, Guanyu Song, Igor Gilitschenski, Marco Pavone, Boris Ivanovic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14521-14530
High-definition (HD) maps have played an integral role in the development of modern autonomous vehicle (AV) stacks albeit with high associated labeling and main

tenance costs. As a result many recent works have proposed methods for estimatin
g HD maps online from sensor data enabling AVs to operate outside of previously-
mapped regions. However current online map estimation approaches are developed i
n isolation of their downstream tasks complicating their integration in AV stack
s. In particular they do not produce uncertainty or confidence estimates. In thi
s work we extend multiple state-of-the-art online map estimation methods to addi
tionally estimate uncertainty and show how this enables more tightly integrating
 online mapping with trajectory forecasting. In doing so we find that incorporat
ing uncertainty yields up to 50% faster training convergence and up to 15% bette
r prediction performance on the real-world nuScenes driving dataset.
****
PerceptionGPT: Effectively Fusing Visual Perception into LLM
Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, Tong Zhang; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
27124-27133
The integration of visual inputs with large language models (LLMs) has led to re
markable advancements in multi-modal capabilities giving rise to vision large la
nguage models (VLLMs). However effectively harnessing LLMs for intricate visual
perception tasks such as detection and segmentation remains a challenge. Convent
ional approaches achieve this by transforming perception signals (e.g. bounding
boxes segmentation masks) into sequences of discrete tokens which struggle with
the precision errors and introduces further complexities for training. In this p
aper we present a novel end-to-end framework named PerceptionGPT which represent
 the perception signals using LLM's dynamic token embedding. Specifically we lev
erage lightweight encoders and decoders to handle the perception signals in LLM'
s embedding space which takes advantage of the representation power of the high-
dimensional token embeddings. Our approach significantly eases the training diff
iculties associated with the discrete representations in prior methods. Furtherm
ore owing to our compact representation the inference speed is also greatly boos
ted. Consequently PerceptionGPT enables accurate flexible and efficient handling
 of complex perception signals. We validate the effectiveness of our approach th
rough extensive experiments. The results demonstrate significant improvements ov
er previous methods with only 4% trainable parameters and less than 25% training
 time.
****
Probabilistic Speech-Driven 3D Facial Motion Synthesis: New Benchmarks Methods a
nd Applications
Karren D. Yang, Anurag Ranjan, Jen-Hao Rick Chang, Raviteja Vemulapalli, Oncel T
uzel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2024, pp. 27294-27303
We consider the task of animating 3D facial geometry from speech signal. Existin
g works are primarily deterministic focusing on learning a one-to-one mapping fr
om speech signal to 3D face meshes on small datasets with limited speakers. Whil
e these models can achieve high-quality lip articulation for speakers in the tra
ining set they are unable to capture the full and diverse distribution of 3D fac
ial motions that accompany speech in the real world. Importantly the relationshi
p between speech and facial motion is one-to-many containing both inter-speaker
and intra-speaker variations and necessitating a probabilistic approach. In this
 paper we identify and address key challenges that have so far limited the devel
opment of probabilistic models: lack of datasets and metrics that are suitable f
or training and evaluating them as well as the difficulty of designing a model t
hat generates diverse results while remaining faithful to a strong conditioning
signal as speech. We first propose large-scale benchmark datasets and metrics su
itable for probabilistic modeling. Then we demonstrate a probabilistic model tha
t achieves both diversity and fidelity to speech outperforming other methods acr
oss the proposed benchmarks. Finally we showcase useful applications of probabil
istic models trained on these large-scale datasets: we can generate diverse spee
ch-driven 3D facial motion that matches unseen speaker styles extracted from ref
erence clips; and our synthetic meshes can be used to improve the performance of
 downstream audio-visual models.

****
LASO: Language-guided Affordance Segmentation on 3D Object
Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, Tat-seng Chua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14251-14260
Segmenting affordance in 3D data is key for bridging perception and action in robots. Existing efforts mostly focus on the visual side and overlook the affordance knowledge from a semantic aspect. This oversight not only limits their generalization to unseen objects but more importantly hinders their synergy with large language models (LLMs) which are excellent task planners that can decompose an overarching command into agent-actionable instructions. With this regard we propose a novel task Language-guided Affordance Segmentation on 3D Object (LASO) which challenges a model to segment a 3D object's part relevant to a given affordance question. To facilitate the task we contribute a dataset comprising 19751 point-question pairs covering 8434 object shapes and 870 expert-crafted questions. As a pioneer solution we further propose PointRefer which highlights an adaptive fusion module to identify target affordance regions at different scales. To ensure a text-aware segmentation we adopt a set of affordance queries conditioned on linguistic cues to generate dynamic kernels. These kernels are further used to convolute with point features and generate a segmentation mask. Comprehensive experiments and analyses validate PointRefer's effectiveness. With these efforts We hope that LASO can steer the direction of 3D affordance guiding it towards enhanced integration with the evolving capabilities of LLMs.
****
Riemannian Multinomial Logistics Regression for SPD Neural Networks
Ziheng Chen, Yue Song, Gaowen Liu, Ramana Rao Kompella, Xiao-Jun Wu, Nicu Sebe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17086-17096
Deep neural networks for learning Symmetric Positive Definite (SPD) matrices are gaining increasing attention in machine learning. Despite the significant progress most existing SPD networks use traditional Euclidean classifiers on an approximated space rather than intrinsic classifiers that accurately capture the geometry of SPD manifolds. Inspired by Hyperbolic Neural Networks (HNNs) we propose Riemannian Multinomial Logistics Regression (RMLR) for the classification layers in SPD networks. We introduce a unified framework for building Riemannian classifiers under the metrics pulled back from the Euclidean space and showcase our framework under the parameterized Log-Euclidean Metric (LEM) and Log-Cholesky Metric (LCM). Besides our framework offers a novel intrinsic explanation for the most popular LogEig classifier in existing SPD networks. The effectiveness of our method is demonstrated in three applications: radar recognition human action recognition and electroencephalography (EEG) classification. The code is available at https://github.com/GitZH-Chen/SPDMLR.git.
****
FreGS: 3D Gaussian Splatting with Progressive Frequency Regularization
Jiahui Zhang, Fangneng Zhan, Muyu Xu, Shijian Lu, Eric Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21424-21433
3D Gaussian splatting has achieved very impressive performance in real-time novel view synthesis. However it often suffers from over-reconstruction during Gaussian densification where high-variance image regions are covered by a few large Gaussians only leading to blur and artifacts in the rendered images. We design a progressive frequency regularization (FreGS) technique to tackle the over-reconstruction issue within the frequency space. Specifically FreGS performs coarse-to-fine Gaussian densification by exploiting low-to-high frequency components that can be easily extracted with low-pass and high-pass filters in the Fourier space. By minimizing the discrepancy between the frequency spectrum of the rendered image and the corresponding ground truth it achieves high-quality Gaussian densification and alleviates the over-reconstruction of Gaussian splatting effectively. Experiments over multiple widely adopted benchmarks (e.g. Mip-NeRF360 Tanks-and-Temples and Deep Blending) show that FreGS achieves superior novel view synth

esis and outperforms the state-of-the-art consistently.
****
Discriminative Sample-Guided and Parameter-Efficient Feature Space Adaptation fo
r Cross-Domain Few-Shot Learning
Rashindrie Perera, Saman Halgamuge; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2024, pp. 23794-23804
In this paper we look at cross-domain few-shot classification which presents the
 challenging task of learning new classes in previously unseen domains with few
labelled examples. Existing methods though somewhat effective encounter several
limitations which we alleviate through two significant improvements. First we in
troduce a lightweight parameter-efficient adaptation strategy to address overfit
ting associated with fine-tuning a large number of parameters on small datasets.
 This strategy employs a linear transformation of pre-trained features significa
ntly reducing the trainable parameter count. Second we replace the traditional n
earest centroid classifier with a discriminative sample-aware loss function enha
ncing the model's sensitivity to the inter- and intra-class variances within the
 training set for improved clustering in feature space. Empirical evaluations on
 the Meta-Dataset benchmark showcase that our approach not only improves accurac
y up to 7.7% and 5.3% on previously seen and unseen datasets respectively but al
so achieves the above performance while being at least 3x more parameter-effici
ent than existing methods establishing a new state-of-the-art in cross-domain fe
w-shot learning. Our code is available at https://github.com/rashindrie/DIPA.
****
What Sketch Explainability Really Means for Downstream Tasks?
Hmrishav Bandyopadhyay, Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain,
 Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 10997-11008
In this paper we explore the unique modality of sketch for explainability emphas
ising the profound impact of human strokes compared to conventional pixel-orient
ed studies. Beyond explanations of network behavior we discern the genuine impli
cations of explainability across diverse downstream sketch-related tasks. We pro
pose a lightweight and portable explainability solution -- a seamless plugin tha
t integrates effortlessly with any pre-trained model eliminating the need for re
-training. Demonstrating its adaptability we present four applications: highly s
tudied retrieval and generation and completely novel assisted drawing and sketch
 adversarial attacks. The centrepiece to our solution is a stroke-level attribut
ion map that takes different forms when linked with downstream tasks. By address
ing the inherent non-differentiability of rasterisation we enable explanations a
t both coarse stroke level (SLA) and partial stroke level (P-SLA) each with its
advantages for specific downstream tasks.
****
Neural Exposure Fusion for High-Dynamic Range Object Detection
Emmanuel Onzon, Maximilian Bömer, Fahim Mannan, Felix Heide; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 17564-17573
Computer vision in unconstrained outdoor scenarios must tackle challenging high
dynamic range (HDR) scenes and rapidly changing illumination conditions. Existin
g methods address this problem with multi-capture HDR sensors and a hardware ima
ge signal processor (ISP) that produces a single fused image as input to a downs
tream neural network. The output of the HDR sensor is a set of low dynamic range
 (LDR) exposures and the fusion in the ISP is performed in image space and typic
ally optimized for human perception on a display. Preferring tonemapped content
with smooth transition regions over detail (and noise) in the resulting image th
is image fusion does typically not preserve all information from the LDR exposur
es that may be essential for downstream computer vision tasks. In this work we d
epart from conventional HDR image fusion and propose a learned task-driven fusio
n in the feature domain. Instead of using a single companded image we introduce
a novel local cross-attention fusion mechanism that exploits semantic features f
rom all exposures learned in an end-to-end fashion with supervision from downstr
eam detection losses. The proposed method outperforms all tested conventional HD

R exposure fusion and auto-exposure methods in challenging automotive HDR scenar
ios.
****
EfficientDreamer: High-Fidelity and Robust 3D Creation via Orthogonal-view Diffu
sion Priors
Zhipeng Hu, Minda Zhao, Chaoyi Zhao, Xinyue Liang, Lincheng Li, Zeng Zhao, Chang
jie Fan, Xiaowei Zhou, Xin Yu; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2024, pp. 4949-4958
While image diffusion models have made significant progress in text-driven 3D co
ntent creation they often fail to accurately capture the intended meaning of tex
t prompts especially for view information. This limitation leads to the Janus pr
oblem where multi-faced 3D models are generated under the guidance of such diffu
sion models. In this paper we propose a robust high-quality 3D content generatio
n pipeline by exploiting orthogonal-view image guidance. First we introduce a no
vel 2D diffusion model that generates an image consisting of four orthogonal-vie
w sub-images based on the given text prompt. Then the 3D content is created usin
g this diffusion model. Notably the generated orthogonal-view image provides str
ong geometric structure priors and thus improves 3D consistency. As a result it
effectively resolves the Janus problem and significantly enhances the quality of
 3D content creation. Additionally we present a 3D synthesis fusion network that
 can further improve the details of the generated 3D contents. Both quantitative
 and qualitative evaluations demonstrate that our method surpasses previous text
-to-3D techniques. Project page: https://efficientdreamer.github.io.
****
HOIAnimator: Generating Text-prompt Human-object Animations using Novel Percepti
ve Diffusion Models
Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao C
hen, Ning Li, Hong Qin; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2024, pp. 811-820
To date the quest to rapidly and effectively produce human-object interaction (H
OI) animations directly from textual descriptions stands at the forefront of com
puter vision research. The underlying challenge demands both a discriminating in
terpretation of language and a comprehensive physics-centric model supporting re
al-world dynamics. To ameliorate this paper advocates HOIAnimator a novel and in
teractive diffusion model with perception ability and also ingeniously crafted t
o revolutionize the animation of complex interactions from linguistic narratives
. The effectiveness of our model is anchored in two ground-breaking innovations:
 (1) Our Perceptive Diffusion Models (PDM) brings together two types of models:
one focused on human movements and the other on objects. This combination allows
 for animations where humans and objects move in concert with each other making
the overall motion more realistic. Additionally we propose a Perceptive Message
Passing (PMP) mechanism to enhance the communication bridging the two models ens
uring that the animations are smooth and unified; (2) We devise an Interaction C
ontact Field (ICF) a sophisticated model that implicitly captures the essence of
 HOIs. Beyond mere predictive contact points the ICF assesses the proximity of h
uman and object to their respective environment informed by a probabilistic dist
ribution of interactions learned throughout the denoising phase. Our comprehensi
ve evaluation showcases HOIanimator's superior ability to produce dynamic contex
t-aware animations that surpass existing benchmarks in text-driven animation syn
thesis.
****
SyncTalk: The Devil is in the Synchronization for Talking Head Synthesis
Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, H
ongyan Liu, Zhaoxin Fan; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2024, pp. 666-676
Achieving high synchronization in the synthesis of realistic speech-driven talki
ng head videos presents a significant challenge. Traditional Generative Adversar
ial Networks (GAN) struggle to maintain consistent facial identity while Neural
Radiance Fields (NeRF) methods although they can address this issue often produc
e mismatched lip movements inadequate facial expressions and unstable head poses

. A lifelike talking head requires synchronized coordination of subject identity
 lip movements facial expressions and head poses. The absence of these synchroni
zations is a fundamental flaw leading to unrealistic and artificial outcomes. To
 address the critical issue of synchronization identified as the "devil" in crea
ting realistic talking heads we introduce SyncTalk. This NeRF-based method effec
tively maintains subject identity enhancing synchronization and realism in talki
ng head synthesis. SyncTalk employs a Face-Sync Controller to align lip movement
s with speech and innovatively uses a 3D facial blendshape model to capture accu
rate facial expressions. Our HeadSync Stabilizer optimizes head poses achieving
more natural head movements. The Portrait-Sync Generator restores hair details a
nd blends the generated head with the torso for a seamless visual experience. Ex
tensive experiments and user studies demonstrate that SyncTalk outperforms state
-of-the-art methods in synchronization and realism. We recommend watching the su
pplementary video: https://ziqiaopeng.github.io/synctalk
****
SFOD: Spiking Fusion Object Detector
Yimeng Fan, Wei Zhang, Changsong Liu, Mingyang Li, Wenrui Lu; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp
. 17191-17200
Event cameras characterized by high temporal resolution high dynamic range low p
ower consumption and high pixel bandwidth offer unique capabilities for object d
etection in specialized contexts. Despite these advantages the inherent sparsity
 and asynchrony of event data pose challenges to existing object detection algor
ithms. Spiking Neural Networks (SNNs) inspired by the way the human brain codes
and processes information offer a potential solution to these difficulties. Howe
ver their performance in object detection using event cameras is limited in curr
ent implementations. In this paper we propose the Spiking Fusion Object Detector
 (SFOD) a simple and efficient approach to SNN-based object detection. Specifica
lly we design a Spiking Fusion Module achieving the first-time fusion of feature
 maps from different scales in SNNs applied to event cameras. Additionally throu
gh integrating our analysis and experiments conducted during the pretraining of
the backbone network on the NCAR dataset we delve deeply into the impact of spik
ing decoding strategies and loss functions on model performance. Thereby we esta
blish state-of-the-art classification results based on SNNs achieving 93.7% accu
racy on the NCAR dataset. Experimental results on the GEN1 detection dataset dem
onstrate that the SFOD achieves a state-of-the-art mAP of 32.1% outperforming ex
isting SNN-based approaches. Our research not only underscores the potential of
SNNs in object detection with event cameras but also propels the advancement of
SNNs. Code is available at https://github.com/yimeng-fan/SFOD.
****
Detector-Free Structure from Motion
Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, Xiaowei
Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2024, pp. 21594-21603
We propose a structure-from-motion framework to recover accurate camera poses an
d point clouds from unordered images. Traditional SfM systems typically rely on
the successful detection of repeatable keypoints across multiple views as the fi
rst step which is difficult for texture-poor scenes and poor keypoint detection
may break down the whole SfM system. We propose a detector-free SfM framework to
 draw benefits from the recent success of detector-free matchers to avoid the ea
rly determination of keypoints while solving the multi-view inconsistency issue
of detector-free matchers. Specifically our framework first reconstructs a coars
e SfM model from quantized detector-free matches. Then it refines the model by a
 novel iterative refinement pipeline which iterates between an attention-based m
ulti-view matching module to refine feature tracks and a geometry refinement mod
ule to improve the reconstruction accuracy. Experiments demonstrate that the pro
posed framework outperforms existing detector-based SfM systems on common benchm
ark datasets. We also collect a texture-poor SfM dataset to demonstrate the capa
bility of our framework to reconstruct texture-poor scenes. Based on this framew
ork we take first place in Image Matching Challenge 2023.

****
CG-HOI: Contact-Guided 3D Human-Object Interaction Generation
Christian Diller, Angela Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19888-19901
We propose CG-HOI the first method to address the task of generating dynamic 3D human-object interactions (HOIs) from text. We model the motion of both human and object in an interdependent fashion as semantically rich human motion rarely happens in isolation without any interactions. Our key insight is that explicitly modeling contact between the human body surface and object geometry can be used as strong proxy guidance both during training and inference. Using this guidance to bridge human and object motion enables generating more realistic and physically plausible interaction sequences where the human body and corresponding object move in a coherent manner. Our method first learns to model human motion object motion and contact in a joint diffusion process inter-correlated through cross-attention. We then leverage this learned contact for guidance during inference to synthesize realistic and coherent HOIs. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences and we show two applications highlighting the capabilities of our method. Conditioned on a given object trajectory we can generate the corresponding human motion without re-training demonstrating strong human-object interdependency learning. Our approach is also flexible and can be applied to static real-world 3D scene scans.
****
Towards Surveillance Video-and-Language Understanding: New Dataset Baselines and Challenges
Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, Zhenzhen Jiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22052-22061
Surveillance videos are important for public security. However current surveillance video tasks mainly focus on classifying and localizing anomalous events. Existing methods are limited to detecting and classifying the predefined events with unsatisfactory semantic understanding although they have obtained considerable performance. To address this issue we propose a new research direction of surveillance video-and-language understanding(VALU) and construct the first multimodal surveillance video dataset. We manually annotate the real-world surveillance dataset UCF-Crime with fine-grained event content and timing. Our newly annotated dataset UCA (UCF-Crime Annotation) contains 23542 sentences with an average length of 20 words and its annotated videos are as long as 110.7 hours. Furthermore we benchmark SOTA models for four multimodal tasks on this newly created dataset which serve as new baselines for surveillance VALU. Through experiments we find that mainstream models used in previously public datasets perform poorly on surveillance video demonstrating new challenges in surveillance VALU. We also conducted experiments on multimodal anomaly detection. These results demonstrate that our multimodal surveillance learning can improve the performance of anomaly detection. All the experiments highlight the necessity of constructing this dataset to advance surveillance AI.
****
AdaRevD: Adaptive Patch Exiting Reversible Decoder Pushes the Limit of Image Deblurring
Xintian Mao, Qingli Li, Yan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25681-25690
Despite the recent progress in enhancing the efficacy of image deblurring the limited decoding capability constrains the upper limit of State-Of-The-Art (SOTA) methods. This paper proposes a pioneering work Adaptive Patch Exiting Reversible Decoder (AdaRevD) to explore their insufficient decoding capability. By inheriting the weights of the well-trained encoder we refactor a reversible decoder which scales up the single-decoder training to multi-decoder training while remaining GPU memory-friendly. Meanwhile we show that our reversible structure gradually disentangles high-level degradation degree and low-level blur pattern (residual of the blur image and its sharp counterpart) from compact degradation represen

tation. Besides due to the spatially-variant motion blur kernels different blur patches have various deblurring difficulties. We further introduce a classifier to learn the degradation degree of image patches enabling them to exit at differ ent sub-decoders for speedup. Experiments show that our AdaRevD pushes the limit of image deblurring e.g. achieving 34.60 dB in PSNR on GoPro dataset.
****
Learning to Remove Wrinkled Transparent Film with Polarized Prior
Jiaqi Tang, Ruizheng Wu, Xiaogang Xu, Sixing Hu, Ying-Cong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24987-24996
In this paper we study a new problem Film Removal (FR) which attempts to remove the interference of wrinkled transparent films and reconstruct the original info rmation under films for industrial recognition systems. We first physically mode l the imaging of industrial materials covered by the film. Considering the specu lar highlight from the film can be effectively recorded by the polarized camera we build a practical dataset with polarization information containing paired dat a with and without transparent film. We aim to remove interference from the film (specular highlights and other degradations) with an end-to-end framework. To l ocate the specular highlight we use an angle estimation network to optimize the polarization angle with the minimized specular highlight. The image with minimiz ed specular highlight is set as a prior for supporting the reconstruction networ k. Based on the prior and the polarized images the reconstruction network can de couple all degradations from the film. Extensive experiments show that our frame work achieves SOTA performance in both image reconstruction and industrial downs tream tasks. Our code will be released at https://github.com/jqtangust/FilmRemov al.
****
OpenEQA: Embodied Question Answering in the Era of Foundation Models
Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mik ael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmes h Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulki t Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chri s Paxton, Alexander Sax, Aravind Rajeswaran; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16488-16498
We present a modern formulation of Embodied Question Answering (EQA) as the task of understanding an environment well enough to answer questions about it in nat ural language. An agent can achieve such an understanding by either drawing upon episodic memory exemplified by agents on smart glasses or by actively exploring the environment as in the case of mobile robots. We accompany our formulation w ith OpenEQA -- the first open-vocabulary benchmark dataset for EQA supporting bo th episodic memory and active exploration use cases. OpenEQA contains over 1600 high-quality human generated questions drawn from over 180 real-world environmen ts. In addition to the dataset we also provide an automatic LLM-powered evaluati on protocol that has excellent correlation with human judgement. Using this data set and evaluation protocol we evaluate several state-of-the-art foundation mode ls including GPT-4V and find that they significantly lag behind human-level perf ormance. Consequently OpenEQA stands out as a straightforward measurable and pra ctically relevant benchmark that poses a considerable challenge to current gener ation of foundation models. We hope this inspires and stimulates future research at the intersection of Embodied AI conversational agents and world models.
****
DreamSalon: A Staged Diffusion Framework for Preserving Identity-Context in Edit able Face Generation
Haonan Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2024, pp. 8589-8598
While large-scale pre-trained text-to-image models can synthesize diverse and hi gh-quality human-centered images novel challenges arise with a nuanced task of " identity fine editing" - precisely modifying specific features of a subject whil e maintaining its inherent identity and context. Existing personalization method s either require time-consuming optimization or learning additional encoders ade

pt in "identity re-contextualization". However they often struggle with detailed and sensitive tasks like human face editing. To address these challenges we introduce DreamSalon a noise-guided staged-editing framework uniquely focusing on detailed image manipulations and identity-context preservation. By discerning editing and boosting stages via the frequency and gradient of predicted noises DreamSalon first performs detailed manipulations on specific features in the editing stage guided by high-frequency information and then employs stochastic denoising in the boosting stage to improve image quality. For more precise editing DreamSalon semantically mixes source and target textual prompts guided by differences in their embedding covariances to direct the model's focus on specific manipulation areas. Our experiments demonstrate DreamSalon's ability to efficiently and faithfully edit fine details on human faces outperforming existing methods both qualitatively and quantitatively.
****
Dispel Darkness for Better Fusion: A Controllable Visual Enhancer based on Cross-modal Conditional Adversarial Learning
Hao Zhang, Linfeng Tang, Xinyu Xiang, Xuhui Zuo, Jiayi Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26487-26496
We propose a controllable visual enhancer named DDBF which is based on cross-modal conditional adversarial learning and aims to dispel darkness and achieve better visible and infrared modalities fusion. Specifically a guided restoration module (GRM) is firstly designed to enhance weakened information in the low-light visible modality. The GRM utilizes the light-invariant high-contrast characteristics of the infrared modality as the central target distribution and constructs a multi-level conditional adversarial sample set to enable continuous controlled brightness enhancement of visible images. Then we develop an information fusion module (IFM) to integrate the advantageous features of the enhanced visible image and the infrared image. Thanks to customized explicit information preservation and hue fidelity constraints the IFM produces visually pleasing results with rich textures significant contrast and vivid colors. The brightened visible image and the final fused image compose the dual output of our DDBF to meet the diverse visual preferences of users. We evaluate DDBF on the public datasets achieving state-of-the-art performances of low-light enhancement and information integration that is available for both day and night scenarios. The experiments also demonstrate that our DDBF is effective in improving decision accuracy for object detection and semantic segmentation. Moreover we offer a user-friendly interface for the convenient application of our model. The code is publicly available at https://github.com/HaoZhang1018/DDBF.
****
Querying as Prompt: Parameter-Efficient Learning for Multimodal Language Model
Tian Liang, Jing Huang, Ming Kong, Luyuan Chen, Qiang Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26855-26865
Recent advancements in language models pre-trained on large-scale corpora have significantly propelled developments in the NLP domain and advanced progress in multimodal tasks. In this paper we propose a Parameter-Efficient multimodal language model learning strategy named QaP (Querying as Prompt). Its core innovation is a novel modality-bridging method that allows a set of modality-specific queries to be input as soft prompts into a frozen pre-trained language model. Specifically we introduce an efficient Text-Conditioned Resampler that is easy to incorporate into the language models which enables adaptive injection of text-related multimodal information at different levels of the model through query learning. This approach effectively bridges multimodal information to the language models while fully leveraging its token fusion and representation potential. We validated our method across four datasets in three distinct multimodal tasks. The results demonstrate that our QaP multimodal language model achieves state-of-the-art performance in various tasks with training only 4.6% parameters.
****
DePT: Decoupled Prompt Tuning

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, Jingkuan Song; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 12924-12933
This work breaks through the Base-New Tradeoff (BNT) dilemma in prompt tuning i.
e. the better the tuned model generalizes to the base (or target) task the worse
 it generalizes to new tasks and vice versa. Specifically through an in-depth an
alysis of the learned features of the base and new tasks we observe that the BNT
 stems from a channel bias issue--the vast majority of feature channels are occu
pied by base-specific knowledge leading to the collapse of task-shared knowledge
 important to new tasks. To address this we propose the Decoupled Prompt Tuning
(DePT) framework which decouples base-specific knowledge from feature channels i
nto an isolated feature space during prompt tuning so as to maximally preserve t
ask-shared knowledge in the original feature space for achieving better zero-sho
t generalization on new tasks. Importantly our DePT is orthogonal to existing pr
ompt tuning approaches and can enhance them with negligible additional computati
onal cost. Extensive experiments on several datasets show the flexibility and ef
fectiveness of DePT.
****
Neural Super-Resolution for Real-time Rendering with Radiance Demodulation
Jia Li, Ziling Chen, Xiaolong Wu, Lu Wang, Beibei Wang, Lei Zhang; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 4357-4367
It is time-consuming to render high-resolution images in applications such as vi
deo games and virtual reality and thus super-resolution technologies become incr
easingly popular for real-time rendering. However it is challenging to preserve
sharp texture details keep the temporal stability and avoid the ghosting artifac
ts in real-time super-resolution rendering. To address this issue we introduce r
adiance demodulation to separate the rendered image or radiance into a lighting
component and a material component considering the fact that the light component
 is smoother than the rendered image so that the high-resolution material compon
ent with detailed textures can be easily obtained. We perform the super-resoluti
on on the lighting component only and re-modulate it with the high-resolution ma
terial component to obtain the final super-resolution image with more texture de
tails. A reliable warping module is proposed by explicitly marking the occluded
regions to avoid the ghosting artifacts. To further enhance the temporal stabili
ty we design a frame-recurrent neural network and a temporal loss to aggregate t
he previous and current frames which can better capture the spatial-temporal con
sistency among reconstructed frames. As a result our method is able to produce t
emporally stable results in real-time rendering with high-quality details even i
n the challenging 4 x4 super-resolution scenarios. Code is available at: \href h
ttps://github.com/Riga2/NSRD https://github.com/Riga2/NSRD .
****
Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction
Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, Xiaogang Jin; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2024, pp. 20331-20341
Implicit neural representation has paved the way for new approaches to dynamic s
cene reconstruction. Nonetheless cutting-edge dynamic neural rendering methods r
ely heavily on these implicit representations which frequently struggle to captu
re the intricate details of objects in the scene. Furthermore implicit methods h
ave difficulty achieving real-time rendering in general dynamic scenes limiting
their use in a variety of tasks. To address the issues we propose a deformable 3
D Gaussians splatting method that reconstructs scenes using 3D Gaussians and lea
rns them in canonical space with a deformation field to model monocular dynamic
scenes. We also introduce an annealing smoothing training mechanism with no extr
a overhead which can mitigate the impact of inaccurate poses on the smoothness o
f time interpolation tasks in real-world scenes. Through a differential Gaussian
 rasterizer the deformable 3D Gaussians not only achieve higher rendering qualit
y but also real-time rendering speed. Experiments show that our method outperfor
ms existing methods significantly in terms of both rendering quality and speed m

aking it well-suited for tasks such as novel-view synthesis time interpolation a
nd real-time rendering.
****
Enhancing 3D Object Detection with 2D Detection-Guided Query Anchors
Haoxuanye Ji, Pengpeng Liang, Erkang Cheng; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21178-21187
Multi-camera-based 3D object detection has made notable progress in the past sev
eral years. However we observe that there are cases (e.g. faraway regions) in wh
ich popular 2D object detectors are more reliable than state-of-the-art 3D detec
tors. In this paper to improve the performance of query-based 3D object detector
s we present a novel query generating approach termed QAF2D which infers 3D quer
y anchors from 2D detection results. A 2D bounding box of an object in an image
is lifted to a set of 3D anchors by associating each sampled point within the bo
x with depth yaw angle and size candidates. Then the validity of each 3D anchor
is verified by comparing its projection in the image with its corresponding 2D b
ox and only valid anchors are kept and used to construct queries. The class info
rmation of the 2D bounding box associated with each query is also utilized to ma
tch the predicted boxes with ground truth for the set-based loss. The image feat
ure extraction backbone is shared between the 3D detector and 2D detector by add
ing a small number of prompt parameters. We integrate QAF2D into three popular q
uery-based 3D object detectors and carry out comprehensive evaluations on the nu
Scenes dataset. The largest improvement that QAF2D can bring about on the nuScen
es validation subset is 2.3% NDS and 2.7% mAP. Code is available at https://gith
ub.com/max-vision/QAF2D.
****
Continual Forgetting for Pre-trained Vision Models
Hongbo Zhao, Bolin Ni, Junsong Fan, Yuxi Wang, Yuntao Chen, Gaofeng Meng, Zhaoxi
ang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2024, pp. 28631-28642
For privacy and security concerns the need to erase unwanted information from pr
e-trained vision models is becoming evident nowadays. In real-world scenarios er
asure requests originate at any time from both users and model owners. These req
uests usually form a sequence. Therefore under such a setting selective informat
ion is expected to be continuously removed from a pre-trained model while mainta
ining the rest. We define this problem as continual forgetting and identify two
key challenges. (i) For unwanted knowledge efficient and effective deleting is c
rucial. (ii) For remaining knowledge the impact brought by the forgetting proced
ure should be minimal. To address them we propose Group Sparse LoRA (GS-LoRA). S
pecifically towards (i) we use LoRA modules to fine-tune the FFN layers in Trans
former blocks for each forgetting task independently and towards (ii) a simple g
roup sparse regularization is adopted enabling automatic selection of specific L
oRA groups and zeroing out the others. GS-LoRA is effective parameter-efficient
data-efficient and easy to implement. We conduct extensive experiments on face r
ecognition object detection and image classification and demonstrate that GS-LoR
A manages to forget specific classes with minimal impact on other classes. Codes
 will be released on https://github.com/bjzhb666/GS-LoRA.
****
Real Acoustic Fields: An Audio-Visual Room Acoustics Dataset and Benchmark
Ziyang Chen, Israel D. Gebru, Christian Richardt, Anurag Kumar, William Laney, A
ndrew Owens, Alexander Richard; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2024, pp. 21886-21896
We present a new dataset called Real Acoustic Fields (RAF) that captures real ac
oustic room data from multiple modalities. The dataset includes high-quality and
 densely captured room impulse response data paired with multi-view images and p
recise 6DoF pose tracking data for sound emitters and listeners in the rooms. We
 used this dataset to evaluate existing methods for novel-view acoustic synthesi
s and impulse response generation which previously relied on synthetic data. In
our evaluation we thoroughly assessed existing audio and audio-visual models aga
inst multiple criteria and proposed settings to enhance their performance on rea
l-world data. We also conducted experiments to investigate the impact of incorpo

rating visual data (i.e. images and depth) into neural acoustic field models. Additionally we demonstrated the effectiveness of a simple sim2real approach where a model is pre-trained with simulated data and fine-tuned with sparse real-world data resulting in significant improvements in the few-shot learning approach. RAF is the first dataset to provide densely captured room acoustic data making it an ideal resource for researchers working on audio and audio-visual neural acoustic field modeling techniques.
****
A Generative Approach for Wikipedia-Scale Visual Entity Recognition
Mathilde Caron, Ahmet Iscen, Alireza Fathi, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17313-17322
In this paper we address web-scale visual entity recognition specifically the task of mapping a given query image to one of the 6 million existing entities in Wikipedia. One way of approaching a problem of such scale is using dual encoder models (e.g. CLIP) where all the entity names and query images are embedded into a unified space paving the way for an approximate kNN search. Alternatively it is also possible to re-purpose a captioning model to directly generate the entity names for a given image. In contrast we introduce a novel Generative Entity Recognition (GER) framework which given an input image learns to auto-regressively decode a semantic and discriminative "code" identifying the target entity. Our experiments demonstrate the efficacy of this GER paradigm showcasing state-of-the-art performance on the challenging OVEN benchmark. GER surpasses strong captioning dual-encoder visual matching and hierarchical classification baselines affirming its advantage in tackling the complexities of web-scale recognition.
****
A Physics-informed Low-rank Deep Neural Network for Blind and Universal Lens Aberration Correction
Jin Gong, Runzhao Yang, Weihang Zhang, Jinli Suo, Qionghai Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24861-24870
High-end lenses although offering high-quality images suffer from both insufficient affordability and bulky design which hamper their applications in low-budget scenarios or on low-payload platforms. A flexible scheme is to tackle the optical aberration of low-end lenses computationally. However it is highly demanded but quite challenging to build a general model capable of handling non-stationary aberrations and covering diverse lenses especially in a blind manner. To address this issue we propose a universal solution by extensively utilizing the physical properties of camera lenses: (i) reducing the complexity of lens aberrations i.e. lens-specific non-stationary blur by warping annual-ring-shaped sub-images into rectangular stripes to transform non-uniform degenerations into a uniform one (ii) building a low-dimensional non-negative orthogonal representation of lens blur kernels to cover diverse lenses; (iii) designing a decoupling network to decompose the input low-quality image into several components degenerated by above kernel bases and applying corresponding pre-trained deconvolution networks to reverse the degeneration. Benefiting from the proper incorporation of lenses' physical properties and unique network design the proposed method achieves superb imaging quality wide applicability for various lenses high running efficiency and is totally free of kernel calibration. These advantages bring great potential for scenarios requiring lightweight high-quality photography.
****
Open-Vocabulary Object 6D Pose Estimation
Jaime Corsetti, Davide Boscaini, Changjae Oh, Andrea Cavallaro, Fabio Poiesi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18071-18080
We introduce the new setting of open-vocabulary object 6D pose estimation in which a textual prompt is used to specify the object of interest. In contrast to existing approaches in our setting (i) the object of interest is specified solely through the textual prompt (ii) no object model (e.g. CAD or video sequence) is required at inference and (iii) the object is imaged from two RGBD viewpoints of

different scenes. To operate in this setting we introduce a novel approach that leverages a Vision-Language Model to segment the object of interest from the scenes and to estimate its relative 6D pose. The key of our approach is a carefully devised strategy to fuse object-level information provided by the prompt with local image features resulting in a feature space that can generalize to novel concepts. We validate our approach on a new benchmark based on two popular datasets REAL275 and Toyota-Light which collectively encompass 34 object instances appearing in four thousand image pairs. The results demonstrate that our approach outperforms both a well-established hand-crafted method and a recent deep learning-based baseline in estimating the relative 6D pose of objects in different scenes. Code and dataset are available at https://jcorsetti.github.io/oryon.
****

Plug and Play Active Learning for Object Detection

Chenhongyi Yang, Lichao Huang, Elliot J. Crowley; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17784-17793

Annotating datasets for object detection is an expensive and time-consuming endeavor. To minimize this burden active learning (AL) techniques are employed to select the most informative samples for annotation within a constrained "annotation budget". Traditional AL strategies typically rely on model uncertainty or sample diversity for query sampling while more advanced methods have focused on developing AL-specific object detector architectures to enhance performance. However these specialized approaches are not readily adaptable to different object detectors due to the significant engineering effort required for integration. To overcome this challenge we introduce Plug and Play Active Learning (PPAL) a simple and effective AL strategy for object detection. PPAL is a two-stage method comprising uncertainty-based and diversity-based sampling phases. In the first stage our Difficulty Calibrated Uncertainty Sampling leverage a category-wise difficulty coefficient that combines both classification and localisation difficulties to re-weight instance uncertainties from which we sample a candidate pool for the subsequent diversity-based sampling. In the second stage we propose Category Conditioned Matching Similarity to better compute the similarities of multi-instance images as ensembles of their instance similarities which is used by the k-Means++ algorithm to sample the final AL queries. PPAL makes no change to model architectures or detector training pipelines; hence it can be easily generalized to different object detectors. We benchmark PPAL on the MS-COCO and Pascal VOC datasets using different detector architectures and show that our method outperforms prior work by a large margin. Code is available at https://github.com/ChenhongyiYang/PPAL
****

Calibrating Multi-modal Representations: A Pursuit of Group Robustness without Annotations

Chenyu You, Yifei Min, Weicheng Dai, Jasjeet S. Sekhon, Lawrence Staib, James S. Duncan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26140-26150

Fine-tuning pre-trained vision-language models like CLIP has yielded success on diverse downstream tasks. However several pain points persist for this paradigm: (i) directly tuning entire pre-trained models becomes both time-intensive and computationally costly. Additionally these tuned models tend to become highly specialized limiting their practicality for real-world deployment; (ii) recent studies indicate that pre-trained vision-language classifiers may overly depend on spurious features -- patterns that correlate with the target in training data but are not related to the true labeling function; and (iii) existing studies on mitigating the reliance on spurious features largely based on the assumption that we can identify such features does not provide definitive assurance for real-world applications. As a piloting study this work focuses on exploring mitigating the reliance on spurious features for CLIP without using any group annotation. To this end we systematically study the existence of spurious correlation on CLIP and CILP+ERM. We first following recent work on Deep Feature Reweighting (DFR) verify that last-layer retraining can greatly improve group robustness on pretrai

ned CLIP. In view of them we advocate a lightweight representation calibration method for fine-tuning CLIP by first generating a calibration set using the pretrained CLIP and then calibrating representations of samples within this set through contrastive learning all without the need for group labels. Extensive experiments and in-depth visualizations on several benchmarks validate the effectiveness of our proposals largely reducing reliance and significantly boosting the model generalization.
****
LiSA: LiDAR Localization with Semantic Awareness
Bochun Yang, Zijun Li, Wen Li, Zhipeng Cai, Chenglu Wen, Yu Zang, Matthias Muller, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15271-15280
LiDAR localization is a fundamental task in robotics and computer vision which estimates the pose of a LiDAR point cloud within a global map. Scene Coordinate Regression (SCR) has demonstrated state-of-the-art performance in this task. In SCR a scene is represented as a neural network which outputs the world coordinates for each point in the input point cloud. However SCR treats all points equally during localization ignoring the fact that not all objects are beneficial for localization. For example dynamic objects and repeating structures often negatively impact SCR. To address this problem we introduce LiSA the first method that incorporates semantic awareness into SCR to boost the localization robustness and accuracy. To avoid extra computation or network parameters during inference we distill the knowledge from a segmentation model to the original SCR network. Experiments show the superior performance of LiSA on standard LiDAR localization benchmarks compared to state-of-the-art methods. Applying knowledge distillation not only preserves high efficiency but also achieves higher localization accuracy than introducing extra semantic segmentation modules. We also analyze the benefit of semantic information for LiDAR localization. Our code is released at https://github.com/Ybchun/LiSA.
****
MMM: Generative Masked Motion Model
Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, Chen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1546-1555
Recent advances in text-to-motion generation using diffusion and autoregressive models have shown promising results. However these models often suffer from a trade-off between real-time performance high fidelity and motion editability. To address this gap we introduce MMM a novel yet simple motion generation paradigm based on Masked Motion Model. MMM consists of two key components: (1) a motion tokenizer that transforms 3D human motion into a sequence of discrete tokens in latent space and (2) a conditional masked motion transformer that learns to predict randomly masked motion tokens conditioned on the pre-computed text tokens. By attending to motion and text tokens in all directions MMM explicitly captures inherent dependency among motion tokens and semantic mapping between motion and text tokens. During inference this allows parallel and iterative decoding of multiple motion tokens that are highly consistent with fine-grained text descriptions therefore simultaneously achieving high-fidelity and high-speed motion generation. In addition MMM has innate motion editability. By simply placing mask tokens in the place that needs editing MMM automatically fills the gaps while guaranteeing smooth transitions between editing and non-editing parts. Extensive experiments on the HumanML3D and KIT-ML datasets demonstrate that MMM surpasses current leading methods in generating high-quality motion (evidenced by superior FID scores of 0.08 and 0.429) while offering advanced editing features such as body-part modification motion in-betweening and the synthesis of long motion sequences. In addition MMM is two orders of magnitude faster on a single mid-range GPU than editable motion diffusion models. Our project page is available at https://exitudio.github.io/MMM-page/.
****
PEGASUS: Personalized Generative 3D Avatars with Composable Attributes
Hyunsoo Cha, Byungjun Kim, Hanbyul Joo; Proceedings of the IEEE/CVF Conference o

n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1072-1081
We present PEGASUS a method for constructing a personalized generative 3D face a vatar from monocular video sources. Our generative 3D avatar enables disentangle d controls to selectively alter the facial attributes (e.g. hair or nose) while preserving the identity. Our approach consists of two stages: synthetic database generation and constructing a personalized generative avatar. We generate a syn thetic video collection of the target identity with varying facial attributes wh ere the videos are synthesized by borrowing the attributes from monocular videos of diverse identities. Then we build a person-specific generative 3D avatar tha t can modify its attributes continuously while preserving its identity. Through extensive experiments we demonstrate that our method of generating a synthetic d atabase and creating a 3D generative avatar is the most effective in preserving identity while achieving high realism. Subsequently we introduce a zero-shot app roach to achieve the same goal of generative modeling more efficiently by levera ging a previously constructed personalized generative model.
****
LMDrive: Closed-Loop End-to-End Driving with Large Language Models
Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L. Waslander, Yu Liu, Hon gsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15120-15130
Despite significant recent progress in the field of autonomous driving modern me thods still struggle and can incur serious accidents when encountering long-tail unforeseen events and challenging urban scenarios. On the one hand large langua ge models (LLM) have shown impressive reasoning capabilities that approach "Arti ficial General Intelligence". On the other hand previous autonomous driving meth ods tend to rely on limited-format inputs (e.g. sensor data and navigation waypo ints) restricting the vehicle's ability to understand language information and i nteract with humans. To this end this paper introduces LMDrive a novel language-guided end-to-end closed-loop autonomous driving framework. LMDrive uniquely pro cesses and integrates multi-modal sensor data with natural language instructions enabling interaction with humans and navigation software in realistic instructi onal settings. To facilitate further research in language-based closed-loop auto nomous driving we also publicly release the corresponding dataset which includes approximately 64K instruction-following data clips and the LangAuto benchmark t hat tests the system's ability to handle complex instructions and challenging dr iving scenarios. Extensive closed-loop experiments are conducted to demonstrate LMDrive's effectiveness. To the best of our knowledge we're the very first work to leverage LLMs for closed-loop end-to-end autonomous driving. Code is availabl e at https://github.com/opendilab/LMDrive
****
MCD: Diverse Large-Scale Multi-Campus Dataset for Robot Perception
Thien-Minh Nguyen, Shenghai Yuan, Thien Hoang Nguyen, Pengyu Yin, Haozhi Cao, Li hua Xie, Maciej Wozniak, Patric Jensfelt, Marko Thiel, Justin Ziegenbein, Noel B lunder; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2024, pp. 22304-22313
Perception plays a crucial role in various robot applications. However existing well-annotated datasets are biased towards autonomous driving scenarios while un labelled SLAM datasets are quickly over-fitted and often lack environment and do main variations. To expand the frontier of these fields we introduce a comprehen sive dataset named MCD (Multi-Campus Dataset) featuring a wide range of sensing modalities high-accuracy ground truth and diverse challenging environments acros s three Eurasian university campuses. MCD comprises both CCS (Classical Cylindri cal Spinning) and NRE (Non-Repetitive Epicyclic) lidars high-quality IMUs (Inert ial Measurement Units) cameras and UWB (Ultra-WideBand) sensors. Furthermore in a pioneering effort we introduce semantic annotations of 29 classes over 59k spa rse NRE lidar scans across three domains thus providing a novel challenge to exi sting semantic segmentation research upon this largely unexplored lidar modality . Finally we propose for the first time to the best of our knowledge continuous-time ground truth based on optimization-based registration of lidar-inertial dat a on large survey-grade prior maps which are also publicly released each several

times the size of existing ones. We conduct a rigorous evaluation of numerous state-of-the-art algorithms on MCD report their performance and highlight the challenges awaiting solutions from the research community.
****
Diff-Plugin: Revitalizing Details for Diffusion-based Low-level Tasks
Yuhao Liu, Zhanghan Ke, Fang Liu, Nanxuan Zhao, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4197-4208
Diffusion models trained on large-scale datasets have achieved remarkable progress in image synthesis. However due to the randomness in the diffusion process they often struggle with handling diverse low-level tasks that require details preservation. To overcome this limitation we present a new Diff-Plugin framework to enable a single pre-trained diffusion model to generate high-fidelity results across a variety of low-level tasks. Specifically we first propose a lightweight Task-Plugin module with a dual branch design to provide task-specific priors guiding the diffusion process in preserving image content. We then propose a Plugin-Selector that can automatically select different Task-Plugins based on the text instruction allowing users to edit images by indicating multiple low-level tasks with natural language. We conduct extensive experiments on 8 low-level vision tasks. The results demonstrate the superiority of Diff-Plugin over existing methods particularly in real-world scenarios. Our ablations further validate that Diff-Plugin is stable schedulable and supports robust training across different dataset sizes.
****
AHIVE: Anatomy-aware Hierarchical Vision Encoding for Interactive Radiology Report Retrieval
Sixing Yan, William K. Cheung, Ivor W. Tsang, Keith Chiu, Terence M. Tong, Ka Chun Cheung, Simon See; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14324-14333
Automatic radiology report generation using deep learning models has been recently explored and found promising. Neural decoders are commonly used for the report generation where irrelevant and unfaithful contents are unavoidable. The retrieval-based approach alleviates the limitation by identifying reports which are relevant to the input to assist the generation. To achieve clinically accurate report retrieval we make reference to clinicians' diagnostic steps of examining a radiology image where anatomical and diagnostic details are typically focused and propose a novel hierarchical visual concept representation called anatomy-aware hierarchical vision encoding (AHIVE). To learn AHIVE we first derive a methodology to extract hierarchical diagnostic descriptions from radiology reports and develop a CLIP-based framework for the model training. Also the hierarchical architecture of AHIVE is designed to support interactive report retrieval so that report revision made at one layer can be propagated to the subsequent ones to trigger other necessary revisions. We conduct extensive experiments and show that AHIVE can outperform the SOTA vision-language retrieval methods in terms of clinical accuracy by a large margin. We provide also a case study to illustrate how it enables interactive report retrieval.
****
CyberDemo: Augmenting Simulated Human Demonstration for Real-World Dexterous Manipulation
Jun Wang, Yuzhe Qin, Kaiming Kuang, Yigit Korkmaz, Akhilan Gurumoorthy, Hao Su, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17952-17963
We introduce CyberDemo a novel approach to robotic imitation learning that leverages simulated human demonstrations for real-world tasks. By incorporating extensive data augmentation in a simulated environment CyberDemo outperforms traditional in-domain real-world demonstrations when transferred to the real world handling diverse physical and visual conditions. Regardless of its affordability and convenience in data collection CyberDemo outperforms baseline methods in terms of success rates across various tasks and exhibits generalizability with previously unseen objects. For example it can rotate novel tetra-valve and penta-valve d

espite human demonstrations only involving tri-valves. Our research demonstrates the significant potential of simulated human demonstrations for real world dexterous manipulation tasks. More details can be found at https://cyber-demo.github.io/
****
MaskCLR: Attention-Guided Contrastive Learning for Robust Action Representation Learning
Mohamed Abdelfattah, Mariam Hassan, Alexandre Alahi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18678-18687
Current transformer-based skeletal action recognition models tend to focus on a limited set of joints and low-level motion patterns to predict action classes. This results in significant performance degradation under small skeleton perturbations or changing the pose estimator between training and testing. In this work we introduce MaskCLR a new Masked Contrastive Learning approach for Robust skeletal action recognition. We propose an Attention-Guided Probabilistic Masking strategy to occlude the most important joints and encourage the model to explore a larger set of discriminative joints. Furthermore we propose a Multi-Level Contrastive Learning paradigm to enforce the representations of standard and occluded skeletons to be class-discriminative i.e. more compact within each class and more dispersed across different classes. Our approach helps the model capture the high-level action semantics instead of low-level joint variations and can be conveniently incorporated into transformer-based models. Without loss of generality we combine MaskCLR with three transformer backbones: the vanilla transformer DSTFormer and STTFormer. Extensive experiments on NTU60 NTU120 and Kinetics400 show that MaskCLR consistently outperforms previous state-of-the-art methods on standard and perturbed skeletons from different pose estimators showing improved accuracy generalization and robustness. Project website: https://maskclr.github.io.
****
Narrative Action Evaluation with Prompt-Guided Multimodal Interaction
Shiyi Zhang, Sule Bai, Guangyi Chen, Lei Chen, Jiwen Lu, Junle Wang, Yansong Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18430-18439
In this paper we investigate a new problem called narrative action evaluation (NAE). NAE aims to generate professional commentary that evaluates the execution of an action. Unlike traditional tasks such as score-based action quality assessment and video captioning involving superficial sentences NAE focuses on creating detailed narratives in natural language. These narratives provide intricate descriptions of actions along with objective evaluations. NAE is a more challenging task because it requires both narrative flexibility and evaluation rigor. One existing possible solution is to use multi-task learning where narrative language and evaluative information are predicted separately. However this approach results in reduced performance for individual tasks because of variations between tasks and differences in modality between language information and evaluation information. To address this we propose a prompt-guided multimodal interaction framework. This framework utilizes a pair of transformers to facilitate the interaction between different modalities of information. It also uses prompts to transform the score regression task into a video-text matching task thus enabling task interactivity. To support further research in this field we re-annotate the MTL-AQA and FineGym datasets with high-quality and comprehensive action narration. Additionally we establish benchmarks for NAE. Extensive experiment results prove that our method outperforms separate learning methods and naive multi-task learning methods. Data and code will be released at https://github.com/shiyi-zh0408/NAE_CVPR2024.
****
R-Cyclic Diffuser: Reductive and Cyclic Latent Diffusion for 3D Clothed Human Digitalization
Kennard Yanting Chan, Fayao Liu, Guosheng Lin, Chuan Sheng Foo, Weisi Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10304-10313

Recently the authors of Zero-1-to-3 demonstrated that a latent diffusion model p
retrained with Internet-scale data can not only address the single-view 3D objec
t reconstruction task but can even attain SOTA results in it. However when appli
ed to the task of single-view 3D clothed human reconstruction Zero-1-to-3 (and r
elated models) are unable to compete with the corresponding SOTA methods in this
 field despite being trained on clothed human data. In this work we aim to tailo
r Zero-1-to-3's approach to the single-view 3D clothed human reconstruction task
 in a much more principled and structured manner. To this end we propose R-Cycli
c Diffuser a framework that adapts Zero-1-to-3's novel approach to clothed human
 data by fusing it with a pixel-aligned implicit model. R-Cyclic Diffuser offers
 a total of three new contributions. The first and primary contribution is R-Cyc
lic Diffuser's cyclical conditioning mechanism for novel view synthesis. This me
chanism directly addresses the view inconsistency problem faced by Zero-1-to-3 a
nd related models. Secondly we further enhance this mechanism with two key featu
res - Lateral Inversion Constraint and Cyclic Noise Selection. Both features are
 designed to regularize and restrict the randomness of outputs generated by a la
tent diffusion model. Thirdly we show how SMPL-X body priors can be incorporated
 in a latent diffusion model such that novel views of clothed human bodies can b
e generated much more accurately. Our experiments show that R-Cyclic Diffuser is
 able to outperform current SOTA methods in single-view 3D clothed human reconst
ruction both qualitatively and quantitatively. Our code is made publicly availab
le at https://github.com/kcyt/r-cyclic-diffuser.
****
Intelligent Grimm - Open-ended Visual Storytelling via Latent Diffusion Models
Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, Weidi Xie; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 6190-6200
Generative models have recently exhibited exceptional capabilities in text-to-im
age generation but still struggle to generate image sequences coherently. In thi
s work we focus on a novel yet challenging task of generating a coherent image s
equence based on a given storyline denoted as open-ended visual storytelling. We
 make the following three contributions: (i) to fulfill the task of visual story
telling we propose a learning-based auto-regressive image generation model terme
d as StoryGen with a novel vision-language context module that enables to genera
te the current frame by conditioning on the corresponding text prompt and preced
ing image-caption pairs; (ii) to address the data shortage of visual storytellin
g we collect paired image-text sequences by sourcing from online videos and open
-source E-books establishing processing pipeline for constructing a large-scale
dataset with diverse characters storylines and artistic styles named StorySalon;
 (iii) Quantitative experiments and human evaluations have validated the superio
rity of our StoryGen where we show it can generalize to unseen characters withou
t any optimization and generate image sequences with coherent content and consis
tent character. Code dataset and models are available at https://haoningwu3639.g
ithub.io/StoryGen_Webpage/
****
Validating Privacy-Preserving Face Recognition under a Minimum Assumption
Hui Zhang, Xingbo Dong, YenLung Lai, Ying Zhou, Xiaoyan Zhang, Xingguo Lv, Zhe J
in, Xuejun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2024, pp. 12205-12214
The widespread use of cloud-based face recognition technology raises privacy con
cerns as unauthorized access to face images can expose personal information or b
e exploited for fraudulent purposes. In response privacy-preserving face recogni
tion (PPFR) schemes have emerged to hide visual information and thwart unauthori
zed access. However the validation methods employed by these schemes often rely
on unrealistic assumptions leaving doubts about their true effectiveness in safe
guarding facial privacy. In this paper we introduce a new approach to privacy va
lidation called Minimum Assumption Privacy Protection Validation (Map^2V). This
is the first exploration of formulating a privacy validation method utilizing de
ep image priors and zeroth-order gradient estimation with the potential to serve
 as a general framework for PPFR evaluation. Building upon Map^2V we comprehensi

vely validate the privacy-preserving capability of PPFRs through a combination o
f human and machine vision. The experiment results and analysis demonstrate the
effectiveness and generalizability of the proposed Map^2V showcasing its superio
rity over native privacy validation methods from PPFR works of literature. Addit
ionally this work exposes privacy vulnerabilities in evaluated state-of-the-art
PPFR schemes laying the foundation for the subsequent effective proposal of coun
termeasures. The source code is available at https://github.com/Beauty9882/MAP2V
.
****

Long-Tailed Anomaly Detection with Learnable Class Names
Chih-Hui Ho, Kuan-Chuan Peng, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12435-12446
Anomaly detection (AD) aims to identify defective images and localize their defe
cts (if any). Ideally AD models should be able to detect defects over many image
 classes; without relying on hard-coded class names that can be uninformative or
 inconsistent across datasets; learn without anomaly supervision; and be robust
to the long-tailed distributions of real-world applications. To address these ch
allenges we formulate the problem of long-tailed AD by introducing several datas
ets with different levels of class imbalance and metrics for performance evaluat
ion. We then propose a novel method LTAD to detect defects from multiple and lon
g-tailed classes without relying on dataset class names. LTAD combines AD by rec
onstruction and semantic AD modules. AD by reconstruction is implemented with a
transformer-based reconstruction module. Semantic AD is implemented with a binar
y classifier which relies on learned pseudo class names and a pretrained foundat
ion model. These modules are learned over two phases. Phase 1 learns the pseudo-
class names and a variational autoencoder (VAE) for feature synthesis that augme
nts the training data to combat long-tails. Phase 2 then learns the parameters o
f the reconstruction and classification modules of LTAD. Extensive experiments u
sing the proposed long-tailed datasets show that LTAD substantially outperforms
the state-of-the-art methods for most forms of dataset imbalance. The long-taile
d dataset split is available at https://zenodo.org/records/10854201
****

ArGue: Attribute-Guided Prompt Tuning for Vision-Language Models
Xinyu Tian, Shu Zou, Zhaoyuan Yang, Jing Zhang; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28578-28587
Although soft prompt tuning is effective in efficiently adapting Vision-Language
 (V&L) models for downstream tasks it shows limitations in dealing with distribu
tion shifts. We address this issue with Attribute-Guided Prompt Tuning (ArGue) m
aking three key contributions. 1) In contrast to the conventional approach of di
rectly appending soft prompts preceding class names we align the model with prim
itive visual attributes generated by Large Language Models (LLMs). We posit that
 a model's ability to express high confidence in these attributes signifies its
capacity to discern the correct class rationales. 2) We introduce attribute samp
ling to eliminate disadvantageous attributes thus only semantically meaningful a
ttributes are preserved. 3) We propose negative prompting explicitly enumerating
 class-agnostic attributes to activate spurious correlations and encourage the m
odel to generate highly orthogonal probability distributions in relation to thes
e negative features. In experiments our method significantly outperforms current
 state-of-the-art prompt tuning methods on both novel class prediction and out-o
f-distribution generalization tasks.
****

Rapid 3D Model Generation with Intuitive 3D Input
Tianrun Chen, Chaotao Ding, Shangzhan Zhang, Chunan Yu, Ying Zang, Zejian Li, Si
da Peng, Lingyun Sun; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2024, pp. 12554-12564
With the emergence of AR/VR 3D models are in tremendous demand. However conventi
onal 3D modeling with Computer-Aided Design software requires much expertise and
 is difficult for novice users. We find that AR/VR devices in addition to servin
g as effective display mediums can offer a promising potential as an intuitive 3
D model creation tool especially with the assistance of AI generative models. He

re we propose Deep3DVRSketch the first 3D model generation network that inputs 3D VR sketches from novice users and generates highly consistent 3D models in multiple categories within seconds irrespective of the users' drawing abilities. We also contribute KO3D+ the largest 3D sketch-shape dataset. Our method pre-trains a conditional diffusion model on quality 3D data then fine-tunes an encoder to map 3D sketches onto the generator's manifold using an adaptive curriculum strategy for limited ground truths. In our experiment our approach achieves state-of-the-art performance in both model quality and fidelity with real-world input from novice users and users can even draw and obtain very detailed geometric structures. In our user study users were able to complete the 3D modeling tasks over 10 times faster using our approach compared to conventional CAD software tools. We believe that our Deep3DVRSketch and KO3D+ dataset can offer a promising solution for future 3D modeling in metaverse era. Check the project page at http://research.kokoni3d.com/Deep3DVRSketch.
****
GenTron: Diffusion Transformers for Image and Video Generation
Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, Juan-Manuel Perez-Rua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6441-6451
In this study we explore Transformer based diffusion models for image and video generation. Despite the dominance of Transformer architectures in various fields due to their flexibility and scalability the visual generative domain primarily utilizes CNN-based U-Net architectures particularly in diffusion-based models. We introduce GenTron a family of Generative models employing Transformer-based diffusion to address this gap. Our initial step was to adapt Diffusion Transformers (DiTs) from class to text conditioning a process involving thorough empirical exploration of the conditioning mechanism. We then scale GenTron from approximately 900M to over 3B parameters observing improvements in visual quality. Furthermore we extend GenTron to text-to-video generation incorporating novel motion-free guidance to enhance video quality. In human evaluations against SDXL GenTron achieves a 51.1% win rate in visual quality (with a 19.8% draw rate) and a 42.3% win rate in text alignment (with a 42.9% draw rate). GenTron notably performs well in T2I-CompBench highlighting its compositional generation ability. We hope GenTron could provide meaningful insights and serve as a valuable reference for future research. Please refer to the arXiv version for the most up-to-date results: https://arxiv.org/abs/2312.04557.
****
Close Imitation of Expert Retouching for Black-and-White Photography
Seunghyun Shin, Jisu Shin, Jihwan Bae, Inwook Shim, Hae-Gon Jeon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25037-25046
Since the widespread availability of cameras black-and-white (BW)photography has been a popular choice for artistic and aesthetic expression. It highlights the main subject in varying tones of gray creating various effects such as drama and contrast. However producing BW photography often demands high-end cameras or photographic editing from experts. Even the experts prefer different styles depending on the subject or even the same subject when taking grayscale photos or converting color images to BW. It is thus questionable which approach is better. To imitate the artistic values of decolorized images this paper introduces a deep metric learning framework with a novel subject-style specified proxy and a large-scale BW dataset. Our proxy-based decolorization utilizes a hierarchical proxy-based loss and a hierarchical bilateral grid network to mimic the experts' retouching scheme. The proxy-based loss captures both expert-discriminative and classsharing characteristics while the hierarchical bilateral grid network enables imitating spatially-variant retouching by considering both global and local scene contexts. Our dataset including color and BW images edited by three experts demonstrates the scalability of our method which can be further enhanced by constructing additional proxies from any set of BW photos like Internet downloaded figures. Our Experiments show that our framework successfully produce visually-pleasing BW images from color ones as evaluated by user preference with respect to arti

stry and aesthetics.
****
TRIP: Temporal Residual Learning with Image Noise Prior for Image-to-Video Diffu
sion Models
Zhongwei Zhang, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Ting Yao, Yang Cao, Tao M
ei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2024, pp. 8671-8681
Recent advances in text-to-video generation have demonstrated the utility of pow
erful diffusion models. Nevertheless the problem is not trivial when shaping dif
fusion models to animate static image (i.e. image-to-video generation). The diff
iculty originates from the aspect that the diffusion process of subsequent anima
ted frames should not only preserve the faithful alignment with the given image
but also pursue temporal coherence among adjacent frames. To alleviate this we p
resent TRIP a new recipe of image-to-video diffusion paradigm that pivots on ima
ge noise prior derived from static image to jointly trigger inter-frame relation
al reasoning and ease the coherent temporal modeling via temporal residual learn
ing. Technically the image noise prior is first attained through one-step backwa
rd diffusion process based on both static image and noised video latent codes. N
ext TRIP executes a residual-like dual-path scheme for noise prediction: 1) a sh
ortcut path that directly takes image noise prior as the reference noise of each
 frame to amplify the alignment between the first frame and subsequent frames; 2
) a residual path that employs 3D-UNet over noised video and static image latent
 codes to enable inter-frame relational reasoning thereby easing the learning of
 the residual noise for each frame. Furthermore both reference and residual nois
e of each frame are dynamically merged via attention mechanism for final video g
eneration. Extensive experiments on WebVid-10M DTDB and MSR-VTT datasets demonst
rate the effectiveness of our TRIP for image-to-video generation. Please see our
 project page at https://trip-i2v.github.io/TRIP/.
****
TexVocab: Texture Vocabulary-conditioned Human Avatars
Yuxiao Liu, Zhe Li, Yebin Liu, Haoqian Wang; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1715-1725
To adequately utilize the available image evidence in multi-view video-based ava
tar modeling we propose TexVocab a novel avatar representation that constructs a
 texture vocabulary and associates body poses with texture maps for animation. G
iven multi-view RGB videos our method initially back-projects all the available
images in the training videos to the posed SMPL surface producing texture maps i
n the SMPL UV domain. Then we construct pairs of human poses and texture maps to
 establish a texture vocabulary for encoding dynamic human appearances under var
ious poses. Unlike the commonly used joint-wise manner we further design a body-
part-wise encoding strategy to learn the structural effects of the kinematic cha
in. Given a driving pose we query the pose feature hierarchically by decomposing
 the pose vector into several body parts and interpolating the texture features
for synthesizing fine-grained human dynamics. Overall our method is able to crea
te animatable human avatars with detailed and dynamic appearances from RGB video
s and the experiments show that our method outperforms state-of-the-art approach
es.
****
KITRO: Refining Human Mesh by 2D Clues and Kinematic-tree Rotation
Fengyuan Yang, Kerui Gu, Angela Yao; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2024, pp. 1052-1061
2D keypoints are commonly used as an additional cue to refine estimated 3D human
 meshes. Current methods optimize the pose and shape parameters with a reproject
ion loss on the provided 2D keypoints. Such an approach while simple and intuiti
ve has limited effectiveness because the optimal solution is hard to find in amb
iguous parameter space and may sacrifice depth. Additionally divergent gradients
 from distal joints complicate and deviate the refinement of proximal joints in
the kinematic chain. To address these we introduce Kinematic-Tree Rotation (KITR
O) a novel mesh refinement strategy that explicitly models depth and human kinem
atic-tree structure. KITRO treats refinement from a bone-wise perspective. Unlik

e previous methods which perform gradient-based optimizations our method calculates bone directions in closed form. By accounting for the 2D pose bone length and parent joint's depth the calculation results in two possible directions for each child joint. We then use a decision tree to trace binary choices for all bones along the human skeleton's kinematic-tree to select the most probable hypothesis. Our experiments across various datasets and baseline models demonstrate that KITRO significantly improves 3D joint estimation accuracy and achieves an ideal 2D fit simultaneously. Our code available at: https://github.com/MartaYang/KITRO.
****
BoQ: A Place is Worth a Bag of Learnable Queries
Amar Ali-bey, Brahim Chaib-draa, Philippe Giguère; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17794-17803
In visual place recognition accurately identifying and matching images of locations under varying environmental conditions and viewpoints remains a significant challenge. In this paper we introduce a new technique called Bag-of-Queries (BoQ) which learns a set of global queries designed to capture universal place-specific attributes. Unlike existing techniques that employ self-attention and generate the queries directly from the input BoQ employ distinct learnable global queries which probe the input features via cross-attention ensuring consistent information aggregation. In addition this technique provides an interpretable attention mechanism and integrates with both CNN and Vision Transformer backbones. The performance of BoQ is demonstrated through extensive experiments on 14 large-scale benchmarks. It consistently outperforms current state-of-the-art techniques including NetVLAD MixVPR and EigenPlaces. Moreover despite being a global retrieval technique (one-stage) BoQ surpasses two-stage retrieval methods such as Patch-NetVLAD TransVPR and R2Former all while being orders of magnitude faster and more efficient. The code and model weights are publicly available at https://github.com/amaralibey/Bag-of-Queries.
****
SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering
Antoine Guédon, Vincent Lepetit; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5354-5363
We propose a method to allow precise and extremely fast mesh extraction from 3D Gaussian Splatting. Gaussian Splatting has recently become very popular as it yields realistic rendering while being significantly faster to train than NeRFs. It is however challenging to extract a mesh from the millions of tiny 3D Gaussians as these Gaussians tend to be unorganized after optimization and no method has been proposed so far. Our first key contribution is a regularization term that encourages the Gaussians to align well with the surface of the scene. We then introduce a method that exploits this alignment to extract a mesh from the Gaussians using Poisson reconstruction which is fast scalable and preserves details in contrast to the Marching Cubes algorithm usually applied to extract meshes from Neural SDFs. Finally we introduce an optional refinement strategy that binds Gaussians to the surface of the mesh and jointly optimizes these Gaussians and the mesh through Gaussian splatting rendering. This enables easy editing sculpting animating and relighting of the Gaussians by manipulating the mesh instead of the Gaussians themselves. Retrieving such an editable mesh for realistic rendering is done within minutes with our method compared to hours with the state-of-the-art method on SDFs while providing a better rendering quality.
****
Understanding and Improving Source-free Domain Adaptation from a Theoretical Perspective
Yu Mitsuzumi, Akisato Kimura, Hisashi Kashima; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28515-28524
Source-free Domain Adaptation (SFDA) is an emerging and challenging research area that addresses the problem of unsupervised domain adaptation (UDA) without source data. Though numerous successful methods have been proposed for SFDA a theor

etical understanding of why these methods work well is still absent. In this pap
er we shed light on the theoretical perspective of existing SFDA methods. Specif
ically we find that SFDA loss functions comprising discriminability and diversit
y losses work in the same way as the training objective in the theory of self-tr
aining based on the expansion assumption which shows the existence of the target
 error bound. This finding brings two novel insights that enable us to build an
improved SFDA method comprising 1) Model Training with Auto-Adjusting Diversity
Constraint and 2) Augmentation Training with Teacher-Student Framework yielding
a better recognition performance. Extensive experiments on three benchmark datas
ets demonstrate the validity of the theoretical analysis and our method.
****
Learning SO(3)-Invariant Semantic Correspondence via Local Shape Transform
Chunghyun Park, Seungwook Kim, Jaesik Park, Minsu Cho; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22978
-22987
Establishing accurate 3D correspondences between shapes stands as a pivotal chal
lenge with profound implications for computer vision and robotics. However exist
ing self-supervised methods for this problem assume perfect input shape alignmen
t restricting their real-world applicability. In this work we introduce a novel
self-supervised Rotation-Invariant 3D correspondence learner with Local Shape Tr
ansform dubbed RIST that learns to establish dense correspondences between shape
s even under challenging intra-class variations and arbitrary orientations. Spec
ifically RIST learns to dynamically formulate an SO(3)-invariant local shape tra
nsform for each point which maps the SO(3)-equivariant global shape descriptor o
f the input shape to a local shape descriptor. These local shape descriptors are
 provided as inputs to our decoder to facilitate point cloud self- and cross-rec
onstruction. Our proposed self-supervised training pipeline encourages semantica
lly corresponding points from different shapes to be mapped to similar local sha
pe descriptors enabling RIST to establish dense point-wise correspondences. RIST
 demonstrates state-of-the-art performances on 3D part label transfer and semant
ic keypoint transfer given arbitrarily rotated point cloud pairs outperforming e
xisting methods by significant margins.
****
GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence
Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, Vincent Lepetit; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2024, pp. 9903-9913
We present GigaPose a fast robust and accurate method for CAD-based novel object
 pose estimation in RGB images. GigaPose first leverages discriminative "templat
es" rendered images of the CAD models to recover the out-of-plane rotation and t
hen uses patch correspondences to estimate the four remaining parameters. Our ap
proach samples templates in only a two-degrees-of-freedom space instead of the u
sual three and matches the input image to the templates using fast nearest-neigh
bor search in feature space results in a speedup factor of 35x compared to the s
tate of the art. Moreover GigaPose is significantly more robust to segmentation
errors. Our extensive evaluation on the seven core datasets of the BOP challenge
 demonstrates that it achieves state-of-the-art accuracy and can be seamlessly i
ntegrated with existing refinement methods. Additionally we show the potential o
f GigaPose with 3D models predicted by recent work on 3D reconstruction from a s
ingle image relaxing the need for CAD models and making 6D pose object estimatio
n much more convenient. Our source code and trained models are publicly availabl
e at https://github.com/nv-nguyen/gigaPose
****
Imagine Before Go: Self-Supervised Generative Map for Object Goal Navigation
Sixian Zhang, Xinyao Yu, Xinhang Song, Xiaohan Wang, Shuqiang Jiang; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 16414-16425
The Object Goal navigation (ObjectNav) task requires the agent to navigate to a
specified target in an unseen environment. Since the environment layout is unkno
wn the agent needs to infer the unknown contextual objects from partially observ

ations thereby deducing the likely location of the target. Previous end-to-end RL methods capture contextual relationships through implicit representations while they lack notion of geometry. Alternatively modular methods construct local maps for recording the observed geometric structure of unseen environment however lacking the reasoning of contextual relation limits the exploration efficiency. In this work we propose the self-supervised generative map (SGM) a modular method that learns the explicit context relation via self-supervised learning. The SGM is trained to leverage both episodic observations and general knowledge to reconstruct the masked pixels of a cropped global map. During navigation the agent maintains an incomplete local semantic map meanwhile the unknown regions of the local map are generated by the pre-trained SGM. Based on the generated map the agent sets the predicted location of the target as the goal and moves towards it. Experiments on Gibson MP3D and HM3D show the effectiveness of our method.
****
Towards Effective Usage of Human-Centric Priors in Diffusion Models for Text-based Human Image Generation
Junyan Wang, Zhenhong Sun, Zhiyu Tan, Xuanbai Chen, Weihua Chen, Hao Li, Cheng Zhang, Yang Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8446-8455
Vanilla text-to-image diffusion models struggle with generating accurate human images commonly resulting in imperfect anatomies such as unnatural postures or disproportionate limbs. Existing methods address this issue mostly by fine-tuning the model with extra images or adding additional controls --- human-centric priors such as pose or depth maps --- during the image generation phase. This paper explores the integration of these human-centric priors directly into the model fine-tuning stage essentially eliminating the need for extra conditions at the inference stage. We realize this idea by proposing a human-centric alignment loss to strengthen human-related information from the textual prompts within the cross-attention maps. To ensure semantic detail richness and human structural accuracy during fine-tuning we introduce scale-aware and step-wise constraints within the diffusion process according to an in-depth analysis of the cross-attention layer. Extensive experiments show that our method largely improves over state-of-the-art text-to-image models to synthesize high-quality human images based on user-written prompts.
****
A Video is Worth 256 Bases: Spatial-Temporal Expectation-Maximization Inversion for Zero-Shot Video Editing
Maomao Li, Yu Li, Tianyu Yang, Yunfei Liu, Dongxu Yue, Zhihui Lin, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7528-7537
This paper presents a video inversion approach for zero-shot video editing which models the input video with low-rank representation during the inversion process. The existing video editing methods usually apply the typical 2D DDIM inversion or naive spatial-temporal DDIM inversion before editing which leverages time-varying representation for each frame to derive noisy latent. Unlike most existing approaches we propose a Spatial-Temporal Expectation-Maximization (STEM) inversion which formulates the dense video feature under an expectation-maximization manner and iteratively estimates a more compact basis set to represent the whole video. Each frame applies the fixed and global representation for inversion which is more friendly for temporal consistency during reconstruction and editing. Extensive qualitative and quantitative experiments demonstrate that our STEM inversion can achieve consistent improvement on two state-of-the-art video editing methods. Project page: https://stem-inv.github.io/page/.
****
HIPTrack: Visual Tracking with Historical Prompts
Wenrui Cai, Qingjie Liu, Yunhong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19258-19267
Trackers that follow Siamese paradigm utilize similarity matching between template and search region features for tracking. Many methods have been explored to enhance tracking performance by incorporating tracking history to better handle s

cenarios involving target appearance variations such as deformation and occlusio
n. However the utilization of historical information in existing methods is insu
fficient and incomprehensive which typically requires repetitive training and in
troduces a large amount of computation. In this paper we show that by providing
a tracker that follows Siamese paradigm with precise and updated historical info
rmation a significant performance improvement can be achieved with completely un
changed parameters. Based on this we propose a historical prompt network that us
es refined historical foreground masks and historical visual features of the tar
get to provide comprehensive and precise prompts for the tracker. We build a nov
el tracker called HIPTrack based on the historical prompt network which achieves
 considerable performance improvements without the need to retrain the entire mo
del. We conduct experiments on seven datasets and experimental results demonstra
te that our method surpasses the current state-of-the-art trackers on LaSOT LaSO
Text GOT-10k and NfS. Furthermore the historical prompt network can seamlessly i
ntegrate as a plug-and-play module into existing trackers providing performance
enhancements. The source code is available at https://github.com/WenRuiCai/HIPTr
ack.
****
URHand: Universal Relightable Hands
Zhaoxi Chen, Gyeongsik Moon, Kaiwen Guo, Chen Cao, Stanislav Pidhorskyi, Tomas S
imon, Rohan Joshi, Yuan Dong, Yichen Xu, Bernardo Pires, He Wen, Lucas Evans, Bo
 Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, Shoou-
I Yu, Javier Romero, Michael Zollhofer, Yaser Sheikh, Ziwei Liu, Shunsuke Saito;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2024, pp. 119-129
Existing photorealistic relightable hand models require extensive identity-speci
fic observations in different views poses and illuminations and face challenges
in generalizing to natural illuminations and novel identities. To bridge this ga
p we present URHand the first universal relightable hand model that generalizes
across viewpoints poses illuminations and identities. Our model allows few-shot
personalization using images captured with a mobile phone and is ready to be pho
torealistically rendered under novel illuminations. To simplify the personalizat
ion process while retaining photorealism we build a powerful universal relightab
le prior based on neural relighting from multi-view images of hands captured in
a light stage with hundreds of identities. The key challenge is scaling the cros
s-identity training while maintaining personalized fidelity and sharp details wi
thout compromising generalization under natural illuminations. To this end we pr
opose a spatially varying linear lighting model as the neural renderer that take
s physics-inspired shading as input feature. By removing non-linear activations
and bias our specifically designed lighting model explicitly keeps the linearity
 of light transport. This enables single-stage training from light-stage data wh
ile generalizing to real-time rendering under arbitrary continuous illuminations
 across diverse identities. In addition we introduce the joint learning of a phy
sically based model and our neural relighting model which further improves fidel
ity and generalization. Extensive experiments show that our approach achieves su
perior performance over existing methods in terms of both quality and generaliza
bility. We also demonstrate quick personalization of URHand from a short phone s
can of an unseen identity.
****
An N-Point Linear Solver for Line and Motion Estimation with Event Cameras
Ling Gao, Daniel Gehrig, Hang Su, Davide Scaramuzza, Laurent Kneip; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 14596-14605
Event cameras respond primarily to edges---formed by strong gradients---and are
thus particularly well-suited for line-based motion estimation. Recent work has
shown that events generated by a single line each satisfy a polynomial constrain
t which describes a manifold in the space-time volume. Multiple such constraints
 can be solved simultaneously to recover the partial linear velocity and line pa
rameters. In this work we show that with a suitable line parametrization this sy
stem of constraints is actually linear in the unknowns which allows us to design

a novel linear solver. Unlike existing solvers our linear solver (i) is fast and numerically stable since it does not rely on expensive root finding (ii) can solve both minimal and overdetermined systems with more than 5 events and (iii) admits the characterization of all degenerate cases and multiple solutions. The found line parameters are singularity-free and have a fixed scale which eliminates the need for auxiliary constraints typically encountered in previous work. To recover the full linear camera velocity we fuse observations from multiple lines with a novel velocity averaging scheme that relies on a geometrically-motivated residual and thus solves the problem more efficiently than previous schemes which minimize an algebraic residual. Extensive experiments in synthetic and real-world settings demonstrate that our method surpasses the previous work in numerical stability and operates over 600 times faster.
****
GenNBV: Generalizable Next-Best-View Policy for Active 3D Reconstruction
Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, Jiangmiao Pang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16436-16445
While recent advances in neural radiance field enable realistic digitization for large-scale scenes the image-capturing process is still time-consuming and labor-intensive. Previous works attempt to automate this process using the Next-Best-View (NBV) policy for active 3D reconstruction. However the existing NBV policies heavily rely on hand-crafted criteria limited action space or per-scene optimized representations. These constraints limit their cross-dataset generalizability. To overcome them we propose GenNBV an end-to-end generalizable NBV policy. Our policy adopts a reinforcement learning (RL)-based framework and extends typical limited action space to 5D free space. It empowers our agent drone to scan from any viewpoint and even interact with unseen geometries during training. To boost the cross-dataset generalizability we also propose a novel multi-source state embedding including geometric semantic and action representations. We establish a benchmark using the Isaac Gym simulator with the Houses3K and OmniObject3D datasets to evaluate this NBV policy. Experiments demonstrate that our policy achieves a 98.26% and 97.12% coverage ratio on unseen building-scale objects from these datasets respectively outperforming prior solutions.
****
Deep-TROJ: An Inference Stage Trojan Insertion Algorithm through Efficient Weight Replacement Attack
Sabbir Ahmed, Ranyang Zhou, Shaahin Angizi, Adnan Siraj Rakin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24810-24819
To insert Trojan into a Deep Neural Network (DNN) the existing attack assumes the attacker can access the victim's training facilities. However a realistic threat model was recently developed by leveraging memory fault to inject Trojans at the inference stage. In this work we develop a novel Trojan attack by adopting a unique memory fault injection technique that can inject bit-flip into the page table of the main memory. In the main memory each weight block consists of a group of weights located at a specific address of a DRAM row. A bit-flip in the page frame number replaces a target weight block of a DNN model with another replacement weight block. To develop a successful Trojan attack leveraging this unique fault model the attacker must solve three key challenges: i) how to identify a minimum set of target weight blocks to be modified? ii) how to identify the corresponding optimal replacement weight block? iii) how to optimize the trigger to maximize the attacker's objective given a target and replacement weight block set? We address them by proposing a novel Deep-TROJ attack algorithm that can identify a minimum set of vulnerable target and corresponding replacement weight blocks while optimizing the trigger at the same time. We evaluate the performance of our proposed Deep-TROJ on CIFAR-10 CIFAR-100 and ImageNet dataset for sixteen different DNN architectures including vision transformers. Proposed Deep-TROJ is the most successful one to date that does not require access to training facilities while successfully bypassing the existing defenses. Our code is available at https://github.com/ML-Security-Research-LAB/Deep-TROJ.

****
Investigating and Mitigating the Side Effects of Noisy Views for Self-Supervised Clustering Algorithms in Practical Multi-View Scenarios
Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, Xiaofeng Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22957-22966
Multi-view clustering (MVC) aims at exploring category structures among multi-view data in self-supervised manners. Multiple views provide more information than single views and thus existing MVC methods can achieve satisfactory performance. However their performance might seriously degenerate when the views are noisy in practical multi-view scenarios. In this paper we formally investigate the drawback of noisy views and then propose a theoretically grounded deep MVC method (namely MVCAN) to address this issue. Specifically we propose a novel MVC objective that enables un-shared parameters and inconsistent clustering predictions across multiple views to reduce the side effects of noisy views. Furthermore a two-level multi-view iterative optimization is designed to generate robust learning targets for refining individual views' representation learning. Theoretical analysis reveals that MVCAN works by achieving the multi-view consistency complementarity and noise robustness. Finally experiments on extensive public datasets demonstrate that MVCAN outperforms state-of-the-art methods and is robust against the existence of noisy views.
****
EvalCrafter: Benchmarking and Evaluating Large Video Generation Models
Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22139-22149
The vision and language generative models have been overgrown in recent years. For video generation various open-sourced models and public-available services have been developed to generate high-quality videos. However these methods often use a few metrics e.g. FVD or IS to evaluate the performance. We argue that it is hard to judge the large conditional generative models from the simple metrics since these models are often trained on very large datasets with multi-aspect abilities. Thus we propose a novel framework and pipeline for exhaustively evaluating the performance of the generated videos. Our approach involves generating a diverse and comprehensive list of 700 prompts for text-to-video generation which is based on an analysis of real-world user data and generated with the assistance of a large language model. Then we evaluate the state-of-the-art video generative models on our carefully designed benchmark in terms of visual qualities content qualities motion qualities and text-video alignment with 17 well-selected objective metrics. To obtain the final leaderboard of the models we further fit a series of coefficients to align the objective metrics to the users' opinions. Based on the proposed human alignment method our final score shows a higher correlation than simply averaging the metrics showing the effectiveness of the proposed evaluation method.
****
SelfOcc: Self-Supervised Vision-Based 3D Occupancy Prediction
Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19946-19956
3D occupancy prediction is an important task for the robustness of vision-centric autonomous driving which aims to predict whether each point is occupied in the surrounding 3D space. Existing methods usually require 3D occupancy labels to produce meaningful results. However it is very laborious to annotate the occupancy status of each voxel. In this paper we propose SelfOcc to explore a self-supervised way to learn 3D occupancy using only video sequences. We first transform the images into the 3D space (e.g. bird's eye view) to obtain 3D representation of the scene. We directly impose constraints on the 3D representations by treating them as signed distance fields. We can then render 2D images of previous and future frames as self-supervision signals to learn the 3D representations. We propose an MVS-embedded strategy to directly optimize the SDF-induced weights with

multiple depth proposals. Our SelfOcc outperforms the previous best method SceneRF by 58.7% using a single frame as input on SemanticKITTI and is the first self-supervised work that produces reasonable 3D occupancy for surround cameras on nuScenes. SelfOcc produces high-quality depth and achieves state-of-the-art results on novel depth synthesis monocular depth estimation and surround-view depth estimation on the SemanticKITTI KITTI-2015 and nuScenes respectively. Code: https://github.com/huang-yh/SelfOcc.
****

SubT-MRS Dataset: Pushing SLAM Towards All-weather Environments

Shibo Zhao, Yuanjun Gao, Tianhao Wu, Damanpreet Singh, Rushan Jiang, Haoxiang Sun, Mansi Sarawata, Yuheng Qiu, Warren Whittaker, Ian Higgins, Yi Du, Shaoshu Su, Can Xu, John Keller, Jay Karhade, Lucas Nogueira, Sourojit Saha, Ji Zhang, Wenshan Wang, Chen Wang, Sebastian Scherer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22647-22657

Simultaneous localization and mapping (SLAM) is a fundamental task for numerous applications such as autonomous navigation and exploration. Despite many SLAM datasets have been released current SLAM solutions still struggle to have sustained and resilient performance. One major issue is the absence of high-quality datasets including diverse all-weather conditions and a reliable metric for assessing robustness. This limitation significantly restricts the scalability and generalizability of SLAM technologies impacting their development validation and deployment. To address this problem we present SubT-MRS an extremely challenging real-world dataset designed to push SLAM towards all-weather environments to pursue the most robust SLAM performance. It contains multi-degraded environments including over 30 diverse scenes such as structureless corridors varying lighting conditions and perceptual obscurants like smoke and dust; multimodal sensors such as LiDAR fisheye camera IMU and thermal camera; and multiple locomotions like aerial legged and wheeled robots. We developed accuracy and robustness evaluation tracks for SLAM and introduced novel robustness metrics. Comprehensive studies are performed revealing new observations challenges and opportunities for future research.
****

Named Entity Driven Zero-Shot Image Manipulation

Zhida Feng, Li Chen, Jing Tian, JiaXiang Liu, Shikun Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9110-9119

We introduced StyleEntity a zero-shot image manipulation model that utilizes named entities as proxies during its training phase. This strategy enables our model to manipulate images using unseen textual descriptions during inference all within a single training phase. Additionally we proposed an inference technique termed Prompt Ensemble Latent Averaging (PELA). PELA averages the manipulation directions derived from various named entities during inference effectively eliminating the noise directions thus achieving stable manipulation. In our experiments StyleEntity exhibited superior performance in a zero-shot setting compared to other methods. The code model weights and datasets is available at https://github.com/feng-zhida/StyleEntity.
****

Relational Matching for Weakly Semi-Supervised Oriented Object Detection

Wenhao Wu, Hau-San Wong, Si Wu, Tianyou Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27800-27810

Oriented object detection has witnessed significant progress in recent years. However the impressive performance of oriented object detectors is at the huge cost of labor-intensive annotations and deteriorates once the annotated data becomes limited. Semi-supervised learning in which sufficient unannotated data are utilized to enhance the base detector is a promising method to address the annotation deficiency problem. Motivated by weakly supervised learning we introduce annotation-efficient point annotations for unannotated images and propose a weakly semi-supervised method for oriented object detection to balance the detection performance and annotation cost. Specifically we propose a Rotation-Modulated Relational Graph Matching method to match relations of proposals centered on annotate

d points between different models to alleviate the ambiguity of point annotation
s in depicting the oriented object. In addition we further propose a Relational
Rank Distribution Matching method to align the rank distribution on classificati
on and regression between different models. Finally to handle the difficult anno
tated points that both models are confused about we introduce weakly supervised
learning to impose positive signals for difficult point-induced clusters to the
base model and focus the base model on the occupancy between the predictions and
 annotated points. We perform extensive experiments on challenging datasets to d
emonstrate the effectiveness of our proposed weakly semi-supervised method in ef
fectively leveraging unannotated data for significant performance improvement.
****
Rethinking the Representation in Federated Unsupervised Learning with Non-IID Da
ta
Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Fengyuan Yu, Huabin Zhu
, Binhui Yao, Tao Wang, Xiaolin Zheng, Yanchao Tan; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22841-22
850
Federated learning achieves effective performance in modeling decentralized data
. In practice client data are not well-labeled which makes it potential for fede
rated unsupervised learning (FUSL) with non-IID data. However the performance of
 existing FUSL methods suffers from insufficient representations i.e. (1) repres
entation collapse entanglement among local and global models and (2) inconsisten
t representation spaces among local models. The former indicates that representa
tion collapse in local model will subsequently impact the global model and other
 local models. The latter means that clients model data representation with inco
nsistent parameters due to the deficiency of supervision signals. In this work w
e propose FedU2 which enhances generating uniform and unified representation in
FUSL with non-IID data. Specifically FedU2 consists of flexible uniform regulari
zer (FUR) and efficient unified aggregator (EUA). FUR in each client avoids repr
esentation collapse via dispersing samples uniformly and EUA in server promotes
unified representation by constraining consistent client model updating. To exte
nsively validate the performance of FedU2 we conduct both cross-device and cross
-silo evaluation experiments on two benchmark datasets i.e. CIFAR10 and CIFAR100
.
****
Distraction is All You Need: Memory-Efficient Image Immunization against Diffusi
on-Based Image Editing
Ling Lo, Cheng Yu Yeo, Hong-Han Shuai, Wen-Huang Cheng; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2446
2-24471
Recent text-to-image (T2I) diffusion models have revolutionized image editing by
 empowering users to control outcomes using natural language. However the ease o
f image manipulation has raised ethical concerns with the potential for maliciou
s use in generating deceptive or harmful content. To address the concerns we pro
pose an image immunization approach named semantic attack to protect our images
from being manipulated by malicious agents using diffusion models. Our approach
focuses on disrupting the semantic understanding of T2I diffusion models regardi
ng specific content. By attacking the cross-attention mechanism that encodes ima
ge features with text messages during editing we distract the model's attention
regarding the content of our concern. Our semantic attack renders the model unce
rtain about the areas to edit resulting in poorly edited images and contradictin
g the malicious editing attempts. In addition by shifting the attack target towa
rds intermediate attention maps from the final generated image our approach subs
tantially diminishes computational burden and alleviates GPU memory constraints
in comparison to previous methods. Moreover we introduce timestep universal grad
ient updating to create timestep-agnostic perturbations effective across differe
nt input noise levels. By treating the full diffusion process as discrete denois
ing timesteps during the attack we achieve equivalent or even superior immunizat
ion efficacy with nearly half the memory consumption of the previous method. Our
 contributions include a practical and effective approach to safeguard images ag

ainst malicious editing and the proposed method offers robust immunization again
st various image inpainting and editing approaches showcasing its potential for
real-world applications.
****
Knowledge-Enhanced Dual-stream Zero-shot Composed Image Retrieval
Yucheng Suo, Fan Ma, Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26951-26962
We study the zero-shot Composed Image Retrieval (ZS-CIR) task which is to retrie
ve the target image given a reference image and a description without training o
n the triplet datasets. Previous works generate pseudo-word tokens by projecting
 the reference image features to the text embedding space. However they focus on
 the global visual representation ignoring the representation of detailed attrib
utes e.g. color object number and layout. To address this challenge we propose a
 Knowledge-Enhanced Dual-stream zero-shot composed image retrieval framework (KE
Ds). KEDs implicitly models the attributes of the reference images by incorporat
ing a database. The database enriches the pseudo-word tokens by providing releva
nt images and captions emphasizing shared attribute information in various aspec
ts. In this way KEDs recognizes the reference image from diverse perspectives. M
oreover KEDs adopts an extra stream that aligns pseudo-word tokens with textual
concepts leveraging pseudo-triplets mined from image-text pairs. The pseudo-word
 tokens generated in this stream are explicitly aligned with fine-grained semant
ics in the text embedding space. Extensive experiments on widely used benchmarks
 i.e. ImageNet-R COCO object Fashion-IQ and CIRR show that KEDs outperforms prev
ious zero-shot composed image retrieval methods. Code is available at https://gi
thub.com/suoych/KEDs.
****
Taming Self-Training for Open-Vocabulary Object Detection
Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, Vijay Kumar B G, Yumin Su
h, Manmohan Chandraker, Dimitris N. Metaxas; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13938-13947
Recent studies have shown promising performance in open-vocabulary object detect
ion (OVD) by utilizing pseudo labels (PLs) from pretrained vision and language m
odels (VLMs). However teacher-student self-training a powerful and widely used p
aradigm to leverage PLs is rarely explored for OVD. This work identifies two cha
llenges of using self-training in OVD: noisy PLs from VLMs and frequent distribu
tion changes of PLs. To address these challenges we propose SAS-Det that tames s
elf-training for OVD from two key perspectives. First we present a split-and-fus
ion (SAF) head that splits a standard detection into an open-branch and a closed
-branch. This design can reduce noisy supervision from pseudo boxes. Moreover th
e two branches learn complementary knowledge from different training data signif
icantly enhancing performance when fused together. Second in our view unlike in
closed-set tasks the PL distributions in OVD are solely determined by the teache
r model. We introduce a periodic update strategy to decrease the number of updat
es to the teacher thereby decreasing the frequency of changes in PL distribution
s which stabilizes the training process. Extensive experiments demonstrate SAS-D
et is both efficient and effective. SAS-Det outperforms recent models of the sam
e scale by a clear margin and achieves 37.4 AP50 and 29.1 APr on novel categorie
s of the COCO and LVIS benchmarks respectively. Code is available at https://git
hub.com/xiaofeng94/SAS-Det.
****
Grounding and Enhancing Grid-based Models for Neural Fields
Zelin Zhao, Fenglei Fan, Wenlong Liao, Junchi Yan; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19425-194
35
Many contemporary studies utilize grid-based models for neural field representat
ion but a systematic analysis of grid-based models is still missing hindering th
e improvement of those models. Therefore this paper introduces a theoretical fra
mework for grid-based models. This framework points out that these models' appro
ximation and generalization behaviors are determined by grid tangent kernels (GT
K) which are intrinsic properties of grid-based models. The proposed framework f

acilitates a consistent and systematic analysis of diverse grid-based models. Furthermore the introduced framework motivates the development of a novel grid-based model named the Multiplicative Fourier Adaptive Grid (MulFAGrid). The numerical analysis demonstrates that MulFAGrid exhibits a lower generalization bound than its predecessors indicating its robust generalization performance. Empirical studies reveal that MulFAGrid achieves state-of-the-art performance in various tasks including 2D image fitting 3D signed distance field (SDF) reconstruction and novel view synthesis demonstrating superior representation ability. The project website is available at https://sites.google.com/view/cvpr24-2034-submission/home.
****
Bilateral Propagation Network for Depth Completion
Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9763-9772
Depth completion aims to derive a dense depth map from sparse depth measurements with a synchronized color image. Current state-of-the-art (SOTA) methods are predominantly propagation-based which work as an iterative refinement on the initial estimated dense depth. However the initial depth estimations mostly result from direct applications of convolutional layers on the sparse depth map. In this paper we present a Bilateral Propagation Network (BP-Net) that propagates depth at the earliest stage to avoid directly convolving on sparse data. Specifically our approach propagates the target depth from nearby depth measurements via a non-linear model whose coefficients are generated through a multi-layer perceptron conditioned on both radiometric difference and spatial distance. By integrating bilateral propagation with multi-modal fusion and depth refinement in a multi-scale framework our BP-Net demonstrates outstanding performance on both indoor and outdoor scenes. It achieves SOTA on the NYUv2 dataset and ranks 1st on the KITTI depth completion benchmark at the time of submission. Experimental results not only show the effectiveness of bilateral propagation but also emphasize the significance of early-stage propagation in contrast to the refinement stage. Our code and trained models will be available on the project page.
****
ESR-NeRF: Emissive Source Reconstruction Using LDR Multi-view Images
Jinseo Jeong, Junseo Koo, Qimeng Zhang, Gunhee Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4598-4609
Existing NeRF-based inverse rendering methods suppose that scenes are exclusively illuminated by distant light sources neglecting the potential influence of emissive sources within a scene. In this work we confront this limitation using LDR multi-view images captured with emissive sources turned on and off. Two key issues must be addressed: 1) ambiguity arising from the limited dynamic range along with unknown lighting details and 2) the expensive computational cost in volume rendering to backtrace the paths leading to final object colors. We present a novel approach ESR-NeRF leveraging neural networks as learnable functions to represent ray-traced fields. By training networks to satisfy light transport segments we regulate outgoing radiances progressively identifying emissive sources while being aware of reflection areas. The results on scenes encompassing emissive sources with various properties demonstrate the superiority of ESR-NeRF in qualitative and quantitative ways. Our approach also extends its applicability to the scenes devoid of emissive sources achieving lower CD metrics on the DTU dataset.
****
Infer from What You Have Seen Before: Temporally-dependent Classifier for Semi-supervised Video Segmentation
Jiafan Zhuang, Zilei Wang, Yixin Zhang, Zhun Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3575-3584
Due to high expense of human labor one major challenge for semantic segmentation in real-world scenarios is the lack of sufficient pixel-level labels which is more serious when processing video data. To exploit unlabeled data for model training semi-supervised learning methods attempt to construct pseudo labels or vari

ous auxiliary constraints as supervision signals. However most of them just proc ess video data as a set of independent images in a per-frame manner. The rich te mporal relationships are ignored which can serve as valuable clues for represent ation learning. Besides this per-frame recognition paradigm is quite different f rom that of humans. Actually benefited from the internal temporal relevance of v ideo data human would wisely use the distinguished semantic concepts in historic al frames to aid the recognition of the current frame. Motivated by this observa tion we propose a novel temporally-dependent classifier (TDC) to mimic the human -like recognition procedure. Comparing to the conventional classifier TDC can gu ide the model to learn a group of temporally-consistent semantic concepts across frames which essentially provides an implicit and effective constraint. We cond uct extensive experiments on Cityscapes and CamVid and the results demonstrate t he superiority of our proposed method to previous state-of-the-art methods. The code is available at https://github.com/jfzhuang/TDC.
****
Unleashing Channel Potential: Space-Frequency Selection Convolution for SAR Obje ct Detection
Ke Li, Di Wang, Zhangyuan Hu, Wenxuan Zhu, Shaofeng Li, Quan Wang; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 17323-17332
Deep Convolutional Neural Networks (DCNNs) have achieved remarkable performance in synthetic aperture radar (SAR) object detection but this comes at the cost of tremendous computational resources partly due to extracting redundant features within a single convolutional layer. Recent works either delve into model compre ssion methods or focus on the carefully-designed lightweight models both of whic h result in performance degradation. In this paper we propose an efficient convo lution module for SAR object detection called SFS-Conv which increases feature d iversity within each convolutional layer through a shunt-perceive-select strateg y. Specifically we shunt input feature maps into space and frequency aspects. Th e former perceives the context of various objects by dynamically adjusting recep tive field while the latter captures abundant frequency variations and textural features via fractional Gabor transformer. To adaptively fuse features from spac e and frequency aspects a parameter-free feature selection module is proposed to ensure that the most representative and distinctive information are preserved. With SFS-Conv we build a lightweight SAR object detection network called SFS-CNe t. Experimental results show that SFS-CNet outperforms state-of-the-art (SoTA) m odels on a series of SAR object detection benchmarks while simultaneously reduci ng both the model size and computational cost.
****
READ: Retrieval-Enhanced Asymmetric Diffusion for Motion Planning
Takeru Oba, Matthew Walter, Norimichi Ukita; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17974-17984
This paper proposes Retrieval-Enhanced Asymmetric Diffusion (READ) for image-bas ed robot motion planning. Given an image of the scene READ retrieves an initial motion from a database of image-motion pairs and uses a diffusion model to refin e the motion for the given scene. Unlike prior retrieval-based diffusion models that require long forward-reverse diffusion paths READ directly diffuses between the source (retrieved) and target motions resulting in an efficient diffusion p ath. A second contribution of READ is its use of asymmetric diffusion whereby it preserves the kinematic feasibility of the generated motion by forward diffusio n in a low-dimensional latent space while achieving high-resolution motion by re verse diffusion in the original task space using cold diffusion. Experimental re sults on various manipulation tasks demonstrate that READ outperforms state-of-t he-art planning methods while ablation studies elucidate the contributions of as ymmetric diffusion.
****
Video Frame Interpolation via Direct Synthesis with the Event-based Reference
Yuhan Liu, Yongjian Deng, Hao Chen, Zhen Yang; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8477-8487
Video Frame Interpolation (VFI) has witnessed a surge in popularity due to its a

bundant downstream applications. Event-based VFI (E-VFI) has recently propelled the advancement of VFI. Thanks to the high temporal resolution benefits event cameras can bridge the informational void present between successive video frames. Most state-of-the-art E-VFI methodologies follow the conventional VFI paradigm which pivots on motion estimation between consecutive frames to generate intermediate frames through a process of warping and refinement. However this reliance engenders a heavy dependency on the quality and consistency of keyframes rendering these methods susceptible to challenges in extreme real-world scenarios such as missing moving objects and severe occlusion dilemmas. This study proposes a novel E-VFI framework that directly synthesize intermediate frames leveraging event-based reference obviating the necessity for explicit motion estimation and substantially enhancing the capacity to handle motion occlusion. Given the sparse and inherently noisy nature of event data we prioritize the reliability of the event-based reference leading to the development of an innovative event-aware reconstruction strategy for accurate reference generation. Besides we implement a bi-directional event-guided alignment from keyframes to the reference using the introduced E-PCD module. Finally a transformer-based decoder is adopted for prediction refinement. Comprehensive experimental evaluations on both synthetic and real-world datasets underscore the superiority of our approach and its potential to execute high-quality VFI tasks.
****
DSL-FIQA: Assessing Facial Image Quality via Dual-Set Degradation Learning and Landmark-Guided Transformer
Wei-Ting Chen, Gurunandan Krishnan, Qiang Gao, Sy-Yen Kuo, Sizhou Ma, Jian Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2931-2941
Generic Face Image Quality Assessment (GFIQA) evaluates the perceptual quality of facial images which is crucial in improving image restoration algorithms and selecting high-quality face images for downstream tasks. We present a novel transformer-based method for GFIQA which is aided by two unique mechanisms. First a novel Dual-Set Degradation Representation Learning (DSL) mechanism uses facial images with both synthetic and real degradations to decouple degradation from content ensuring generalizability to real-world scenarios. This self-supervised method learns degradation features on a global scale providing a robust alternative to conventional methods that use local patch information in degradation learning. Second our transformer leverages facial landmarks to emphasize visually salient parts of a face image in evaluating its perceptual quality. We also introduce a balanced and diverse Comprehensive Generic Face IQA (CGFIQA-40k) dataset of 40K images carefully designed to overcome the biases in particular the imbalances in skin tone and gender representation in existing datasets. Extensive analysis and evaluation demonstrate the robustness of our method marking a significant improvement over prior methods.
****
FMA-Net: Flow-Guided Dynamic Filtering and Iterative Feature Refinement with Multi-Attention for Joint Video Super-Resolution and Deblurring
Geunhyuk Youk, Jihyong Oh, Munchurl Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 44-55
We present a joint learning scheme of video super-resolution and deblurring called VSRDB to restore clean high-resolution (HR) videos from blurry low-resolution (LR) ones. This joint restoration problem has drawn much less attention compared to single restoration problems. In this paper we propose a novel flow-guided dynamic filtering (FGDF) and iterative feature refinement with multi-attention (FRMA) which constitutes our VSRDB framework denoted as FMA-Net. Specifically our proposed FGDF enables precise estimation of both spatio-temporally-variant degradation and restoration kernels that are aware of motion trajectories through sophisticated motion representation learning. Compared to conventional dynamic filtering the FGDF enables the FMA-Net to effectively handle large motions into the VSRDB. Additionally the stacked FRMA blocks trained with our novel temporal anchor (TA) loss which temporally anchors and sharpens features refine features in a coarse-to-fine manner through iterative updates. Extensive experiments demonstr

ate the superiority of the proposed FMA-Net over state-of-the-art methods in terms of both quantitative and qualitative quality. Codes and pre-trained models are available at: https://kaist-viclab.github.io/fmanet-site.
****
OVMR: Open-Vocabulary Recognition with Multi-Modal References
Zehong Ma, Shiliang Zhang, Longhui Wei, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16571-16581
The challenge of open-vocabulary recognition lies in the model has no clue of new categories it is applied to. Existing works have proposed different methods to embed category cues into the model e.g. through few-shot fine-tuning providing category names or textual descriptions to Vision-Language Models. Fine-tuning is time-consuming and degrades the generalization capability. Textual descriptions could be ambiguous and fail to depict visual details. This paper tackles open-vocabulary recognition from a different perspective by referring to multi-modal clues composed of textual descriptions and exemplar images. Our method named OVMR adopts two innovative components to pursue a more robust category cues embedding. A multi-modal classifier is first generated by dynamically complementing textual descriptions with image exemplars. A preference-based refinement module is hence applied to fuse uni-modal and multi-modal classifiers with the aim to alleviate issues of low-quality exemplar images or textual descriptions. The proposed OVMR is a plug-and-play module and works well with exemplar images randomly crawled from the Internet. Extensive experiments have demonstrated the promising performance of OVMR e.g. it outperforms existing methods across various scenarios and setups. Codes are publicly available at \href https://github.com/Zehong-Ma/OVMR https://github.com/Zehong-Ma/OVMR .
****
Hourglass Tokenizer for Efficient Transformer-Based 3D Human Pose Estimation
Wenhao Li, Mengyuan Liu, Hong Liu, Pichao Wang, Jialun Cai, Nicu Sebe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 604-613
Transformers have been successfully applied in the field of video-based 3D human pose estimation. However the high computational costs of these video pose transformers (VPTs) make them impractical on resource-constrained devices. In this paper we present a plug-and-play pruning-and-recovering framework called Hourglass Tokenizer (HoT) for efficient transformer-based 3D human pose estimation from videos. Our HoT begins with pruning pose tokens of redundant frames and ends with recovering full-length tokens resulting in a few pose tokens in the intermediate transformer blocks and thus improving the model efficiency. To effectively achieve this we propose a token pruning cluster (TPC) that dynamically selects a few representative tokens with high semantic diversity while eliminating the redundancy of video frames. In addition we develop a token recovering attention (TRA) to restore the detailed spatio-temporal information based on the selected tokens thereby expanding the network output to the original full-length temporal resolution for fast inference. Extensive experiments on two benchmark datasets (i.e. Human3.6M and MPI-INF-3DHP) demonstrate that our method can achieve both high efficiency and estimation accuracy compared to the original VPT models. For instance applying to MotionBERT and MixSTE on Human3.6M our HoT can save nearly 50% FLOPs without sacrificing accuracy and nearly 40% FLOPs with only 0.2% accuracy drop respectively. Code and models are available at https://github.com/NationalGAILab/HoT.
****
Boosting Diffusion Models with Moving Average Sampling in Frequency Domain
Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8911-8920
Diffusion models have recently brought a powerful revolution in image generation. Despite showing impressive generative capabilities most of these models rely on the current sample to denoise the next one possibly resulting in denoising instability. In this paper we reinterpret the iterative denoising process as model optimization and leverage a moving average mechanism to ensemble all the prior s

amples. Instead of simply applying moving average to the denoised samples at dif ferent timesteps we first map the denoised samples to data space and then perfor m moving average to avoid distribution shift across timesteps. In view that diff usion models evolve the recovery from low-frequency components to high-frequency details we further decompose the samples into different frequency components an d execute moving average separately on each component. We name the complete appr oach "Moving Average Sampling in Frequency domain (MASF)". MASF could be seamles sly integrated into mainstream pre-trained diffusion models and sampling schedul es. Extensive experiments on both unconditional and conditional diffusion models demonstrate that our MASF leads to superior performances compared to the baseli nes with almost negligible additional complexity cost.
****
GART: Gaussian Articulated Template Models
Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, Kostas Daniilidis; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2024, pp. 19876-19887
We introduce Gaussian Articulated Template Model (GART) an explicit efficient an d expressive representation for non-rigid articulated subject capturing and rend ering from monocular videos. GART utilizes a mixture of moving 3D Gaussians to e xplicitly approximate a deformable subject's geometry and appearance. It takes a dvantage of a categorical template model prior (SMPL SMAL etc.) with learnable f orward skinning while further generalizing to more complex non-rigid deformation s with novel latent bones. GART can be reconstructed via differentiable renderin g from monocular videos in seconds or minutes and rendered in novel poses faster than 150fps.
****
Global and Local Prompts Cooperation via Optimal Transport for Federated Learnin g
Hongxia Li, Wei Huang, Jingya Wang, Ye Shi; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12151-12161
Prompt learning in pretrained visual-language models has shown remarkable flexib ility across various downstream tasks. Leveraging its inherent lightweight natur e recent research attempted to integrate the powerful pretrained models into fed erated learning frameworks to simultaneously reduce communication costs and prom ote local training on insufficient data. Despite these efforts current federated prompt learning methods lack specialized designs to systematically address seve re data heterogeneities e.g. data distribution with both label and feature shift s involved. To address this challenge we present Federated Prompts Cooperation v ia Optimal Transport (FedOTP) which introduces efficient collaborative prompt le arning strategies to capture diverse category traits on a per-client basis. Spec ifically for each client we learn a global prompt to extract consensus knowledge among clients and a local prompt to capture client-specific category characteri stics. Unbalanced Optimal Transport is then employed to align local visual featu res with these prompts striking a balance between global consensus and local per sonalization. By relaxing one of the equality constraints FedOTP enables prompts to focus solely on core image patch regions. Extensive experiments on datasets with various types of heterogeneities have demonstrated that our FedOTP outperfo rms the state-of-the-art methods.
****
Bi-Causal: Group Activity Recognition via Bidirectional Causality
Youliang Zhang, Wenxuan Liu, Danni Xu, Zhuo Zhou, Zheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp . 1450-1459
Current approaches in Group Activity Recognition (GAR) predominantly emphasize H uman Relations (HRs) while often neglecting the impact of Human-Object Interacti ons (HOIs). This study prioritizes the consideration of both HRs and HOIs emphas izing their interdependence. Notably employing Granger Causality Tests reveals t he presence of bidirectional causality between HRs and HOIs. Leveraging this ins ight we propose a Bidirectional-Causal GAR network. This network establishes a c ausality communication channel while modeling relations and interactions enablin

g reciprocal enhancement between human-object interactions and human relations e nsuring their mutual consistency. Additionally an Interaction Module is devised to effectively capture the dynamic nature of human-object interactions. Comprehe nsive experiments conducted on two publicly available datasets showcase the supe riority of our proposed method over state-of-the-art approaches.
****
Space-Time Diffusion Features for Zero-Shot Text-Driven Motion Transfer
Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, Tali Dekel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 24, pp. 8466-8476
We present a new method for text-driven motion transfer - synthesizing a video t hat complies with an input text prompt describing the target objects and scene w hile maintaining an input video's motion and scene layout. Prior methods are con fined to transferring motion across two subjects within the same or closely rela ted object categories and are applicable for limited domains (e.g. humans). In t his work we consider a significantly more challenging setting in which the targe t and source objects differ drastically in shape and fine-grained motion charact eristics (e.g. translating a jumping dog into a dolphin). To this end we leverag e a pre-trained and fixed text-to-video diffusion model which provides us with g enerative and motion priors. The pillar of our method is a new space-time featur e loss derived directly from the model. This loss guides the generation process to preserve the overall motion of the input video while complying with the targe t object in terms of shape and fine-grained motion traits.
****
KP-RED: Exploiting Semantic Keypoints for Joint 3D Shape Retrieval and Deformati on
Ruida Zhang, Chenyangguang Zhang, Yan Di, Fabian Manhardt, Xingyu Liu, Federico Tombari, Xiangyang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20540-20550
In this paper we present KP-RED a unified KeyPoint-driven REtrieval and Deformat ion framework that takes object scans as input and jointly retrieves and deforms the most geometrically similar CAD models from a pre-processed database to tigh tly match the target. Unlike existing dense matching based methods that typicall y struggle with noisy partial scans we propose to leverage category-consistent s parse keypoints to naturally handle both full and partial object scans. Specific ally we first employ a lightweight retrieval module to establish a keypoint-base d embedding space measuring the similarity among objects by dynamically aggregat ing deformation-aware local-global features around extracted keypoints. Objects that are close in the embedding space are considered similar in geometry. Then w e introduce the neural cage-based deformation module that estimates the influenc e vector of each keypoint upon cage vertices inside its local support region to control the deformation of the retrieved shape. Extensive experiments on the syn thetic dataset PartNet and the real-world dataset Scan2CAD demonstrate that KP-R ED surpasses existing state-of-the-art approaches by a large margin. Codes and t rained models will be released in https://github.com/lolrudy/KP-RED.
****
Learning from One Continuous Video Stream
João Carreira, Michael King, Viorica Patraucean, Dilara Gokay, Catalin Ionescu, Yi Yang, Daniel Zoran, Joseph Heyward, Carl Doersch, Yusuf Aytar, Dima Damen, An drew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2024, pp. 28751-28761
We introduce a framework for online learning from a single continuous video stre am - the way people and animals learn without mini-batches data augmentation or shuffling. This poses great challenges given the high correlation between consec utive video frames and there is very little prior work on it. Our framework allo ws us to do a first deep dive into the topic and includes a collection of stream s and tasks composed from two existing video datasets plus methodology for perfo rmance evaluation that considers both adaptation and generalization. We employ p ixel-to-pixel modelling as a practical and flexible way to switch between pre-tr aining and single-stream evaluation as well as between arbitrary tasks without e

ver requiring changes to models and always using the same pixel loss. Equipped with this framework we obtained large single-stream learning gains from pre-training with a novel family of future prediction tasks found that momentum hurts and that the pace of weight updates matters. The combination of these insights leads to matching the performance of IID learning with batch size 1 when using the same architecture and without costly replay buffers.
****
VGGSfM: Visual Geometry Grounded Deep Structure From Motion
Jianyuan Wang, Nikita Karaev, Christian Rupprecht, David Novotny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21686-21697
Structure-from-motion (SfM) is a long-standing problem in the computer vision community which aims to reconstruct the camera poses and 3D structure of a scene from a set of unconstrained 2D images. Classical frameworks solve this problem in an incremental manner by detecting and matching keypoints registering images triangulating 3D points and conducting bundle adjustment. Recent research efforts have predominantly revolved around harnessing the power of deep learning techniques to enhance specific elements (e.g. keypoint matching) but are still based on the original non-differentiable pipeline. Instead we propose a new deep SfM pipeline where each component is fully differentiable and thus can be trained in an end-to-end manner. To this end we introduce new mechanisms and simplifications. First we build on recent advances in deep 2D point tracking to extract reliable pixel-accurate tracks which eliminates the need for chaining pairwise matches. Furthermore we recover all cameras simultaneously based on the image and track features instead of gradually registering cameras. Finally we optimise the cameras and triangulate 3D points via a differentiable bundle adjustment layer. We attain state-of-the-art performance on three popular datasets CO3D IMC Phototourism and ETH3D.
****
MIGC: Multi-Instance Generation Controller for Text-to-Image Synthesis
Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6818-6828
We present a Multi-Instance Generation (MIG) task simultaneously generating multiple instances with diverse controls in one image. Given a set of predefined coordinates and their corresponding descriptions the task is to ensure that generated instances are accurately at the designated locations and that all instances' attributes adhere to their corresponding description. This broadens the scope of current research on Single-instance generation elevating it to a more versatile and practical dimension. Inspired by the idea of divide and conquer we introduce an innovative approach named Multi-Instance Generation Controller (MIGC) to address the challenges of the MIG task. Initially we break down the MIG task into several subtasks each involving the shading of a single instance. To ensure precise shading for each instance we introduce an instance enhancement attention mechanism. Lastly we aggregate all the shaded instances to provide the necessary information for accurately generating multiple instances in stable diffusion (SD). To evaluate how well generation models perform on the MIG task we provide a COCO-MIG benchmark along with an evaluation pipeline. Extensive experiments were conducted on the proposed COCO-MIG benchmark as well as on various commonly used benchmarks. The evaluation results illustrate the exceptional control capabilities of our model in terms of quantity position attribute and interaction. Code and demos will be released at https://migcproject.github.io/.
****
Distilling CLIP with Dual Guidance for Learning Discriminative Human Body Shape Representation
Feng Liu, Minchul Kim, Zhiyuan Ren, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 256-266
Person Re-Identification (ReID) holds critical importance in computer vision with pivotal applications in public safety and crime prevention. Traditional ReID methods reliant on appearance attributes such as clothing and color encounter lim

itations in long-term scenarios and dynamic environments. To address these chall enges we propose CLIP3DReID an innovative approach that enhances person ReID by integrating linguistic descriptions with visual perception leveraging pretrained CLIP model for knowledge distillation. Our method first employs CLIP to automat ically label body shapes with linguistic descriptors. We then apply optimal tran sport theory to align the student model's local visual features with shape-aware tokens derived from CLIP's linguistic output. Additionally we align the student model's global visual features with those from the CLIP image encoder and the 3 D SMPL identity space fostering enhanced domain robustness. CLIP3DReID notably e xcels in discerning discriminative body shape features achieving state-of-the-ar t results in person ReID. Our approach represents a significant advancement in R eID offering robust solutions to existing challenges and setting new directions for future research.
****
Retrieval-Augmented Open-Vocabulary Object Detection
Jooyeon Kim, Eulrang Cho, Sehyung Kim, Hyunwoo J. Kim; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17427 -17436
Open-vocabulary object detection (OVD) has been studied with Vision-Language Mod els (VLMs) to detect novel objects beyond the pre-trained categories. Previous a pproaches improve the generalization ability to expand the knowledge of the dete ctor using 'positive' pseudo-labels with additional 'class' names e.g. sock iPod and alligator. To extend the previous methods in two aspects we propose Retriev al-Augmented Losses and visual Features (RALF). Our method retrieves related 'ne gative' classes and augments loss functions. Also visual features are augmented with 'verbalized concepts' of classes e.g. worn on the feet handheld music playe r and sharp teeth. Specifically RALF consists of two modules: Retrieval Augmente d Losses (RAL) and Retrieval-Augmented visual Features (RAF). RAL constitutes tw o losses reflecting the semantic similarity with negative vocabularies. In addit ion RAF augments visual features with the verbalized concepts from a large langu age model (LLM). Our experiments demonstrate the effectiveness of RALF on COCO a nd LVIS benchmark datasets. We achieve improvement up to 3.4 box AP_ 50 ^ \text N   on novel categories of the COCO dataset and 3.6 mask AP_ \text r   gains on the LVIS dataset. Code is available at https://github.com/mlvlab/RALF.
****
MULTIFLOW: Shifting Towards Task-Agnostic Vision-Language Pruning
Matteo Farina, Massimiliano Mancini, Elia Cunegatti, Gaowen Liu, Giovanni Iacca, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2024, pp. 16185-16195
While excellent in transfer learning Vision-Language models (VLMs) come with hig h computational costs due to their large number of parameters. To address this i ssue removing parameters via model pruning is a viable solution. However existin g techniques for VLMs are task-specific and thus require pruning the network fro m scratch for each new task of interest. In this work we explore a new direction : Task-Agnostic Vision-Language Pruning (TA-VLP). Given a pretrained VLM the goa l is to find a unique pruned counterpart transferable to multiple unknown downst ream tasks. In this challenging setting the transferable representations already encoded in the pretrained model are a key aspect to preserve. Thus we propose M ultimodal Flow Pruning (MULTIFLOW) a first gradient-free pruning framework for T A-VLP where: (i) the importance of a parameter is expressed in terms of its magn itude and its information flow by incorporating the saliency of the neurons it c onnects; and (ii) pruning is driven by the emergent (multimodal) distribution of the VLM parameters after pretraining. We benchmark eight state-of-the-art pruni ng algorithms in the context of TA-VLP experimenting with two VLMs three vision-language tasks and three pruning ratios. Our experimental results show that MULT IFLOW outperforms recent sophisticated combinatorial competitors in the vast maj ority of the cases paving the way towards addressing TA-VLP. The code is publicl y available at https://github.com/FarinaMatteo/multiflow.
****
Spin-UP: Spin Light for Natural Light Uncalibrated Photometric Stereo

Zongrui Li, Zhan Lu, Haojie Yan, Boxin Shi, Gang Pan, Qian Zheng, Xudong Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11905-11914

Natural Light Uncalibrated Photometric Stereo (NaUPS) relieves the strict environment and light assumptions in classical Uncalibrated Photometric Stereo (UPS) methods. However due to the intrinsic ill-posedness and high-dimensional ambiguities addressing NaUPS is still an open question. Existing works impose strong assumptions on the environment lights and objects' material restricting the effectiveness in more general scenarios. Alternatively some methods leverage supervised learning with intricate models while lacking interpretability resulting in a biased estimation. In this work we proposed Spin Light Uncalibrated Photometric Stereo (Spin-UP) an unsupervised method to tackle NaUPS in various environment lights and objects. The proposed method uses a novel setup that captures the object's images on a rotatable platform which mitigates NaUPS's ill-posedness by reducing unknowns and provides reliable priors to alleviate NaUPS's ambiguities. Leveraging neural inverse rendering and the proposed training strategies Spin-UP recovers surface normals environment light and isotropic reflectance under complex natural light with low computational cost. Experiments have shown that Spin-UP outperforms other supervised / unsupervised NaUPS methods and achieves state-of-the-art performance on synthetic and real-world datasets. Codes and data are available at https://github.com/LMozart/CVPR2024-SpinUP.

****

LLaFS: When Large Language Models Meet Few-Shot Segmentation

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3065-3075

This paper proposes LLaFS the first attempt to leverage large language models (LLMs) in few-shot segmentation. In contrast to the conventional few-shot segmentation methods that only rely on the limited and biased information from the annotated support images LLaFS leverages the vast prior knowledge gained by LLM as an effective supplement and directly uses the LLM to segment images in a few-shot manner. To enable the text-based LLM to handle image-related tasks we carefully design an input instruction that allows the LLM to produce segmentation results represented as polygons and propose a region-attribute table to simulate the human visual mechanism and provide multi-modal guidance. We also synthesize pseudo samples and use curriculum learning for pretraining to augment data and achieve better optimization. LLaFS achieves state-of-the-art results on multiple datasets showing the potential of using LLMs for few-shot computer vision tasks.

****

Kernel Adaptive Convolution for Scene Text Detection via Distance Map Prediction

Jinzhi Zheng, Heng Fan, Libo Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5957-5966

Segmentation-based scene text detection algorithms that are accurate to the pixel level can satisfy the detection of arbitrary shape scene text and have received widespread attention. On the one hand due to the complexity and diversity of the scene text the convolution with a fixed kernel size has some limitations in extracting the visual features of the scene text. On the other hand most of the existing segmentation-based algorithms only segment the center of the text losing information such as the edges and directions of the text with limited detection accuracy. There are also some improved algorithms that use iterative corrections or introduce other multiple information to improve text detection accuracy but at the expense of efficiency. To address these issues this paper proposes a simple and effective scene text detection method the Kernel Adaptive Convolution which is designed with a Kernel Adaptive Convolution Module for scene text detection via predicting the distance map. Specifically first we design an extensible kernel adaptive convolution module (KACM) to extract visual features from multiple convolutions with different kernel sizes in an adaptive manner. Secondly our method predicts the text distance map under the supervision of a priori information (including direction map and foreground segmentation map) and completes the text detection from the predicted distance map. Experiments on four publicly avai

lable datasets prove the effectiveness of our algorithm in which the accuracy and efficiency of both the Total-Text and TD500 outperform the state-of-the-art algorithm. The algorithm efficiency is improved while the accuracy is competitive on ArT and CTW1500.
****
PixelLM: Pixel Reasoning with Large Multimodal Model
Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, Xiaojie Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26374-26383
While large multimodal models (LMMs) have achieved remarkable progress generating pixel-level masks for image reasoning tasks involving multiple open-world targets remains a challenge. To bridge this gap we introduce PixelLM an effective and efficient LMM for pixel-level reasoning and understanding. Central to PixelLM are a novel lightweight pixel decoder and a comprehensive segmentation codebook. The decoder efficiently produces masks from the hidden embeddings of the codebook tokens which encode detailed target-relevant information. With this design PixelLM harmonizes with the structure of popular LMMs and avoids the need for additional costly segmentation models. Furthermore we propose a token fusion method to enhance the model's ability to differentiate between multiple targets leading to substantially improved mask quality. To advance research in this area we construct MUSE a high-quality multi-target reasoning segmentation benchmark. PixelLM excels across various pixel-level image reasoning and understanding tasks outperforming well-established methods in multiple benchmarks including MUSE and multi-referring segmentation. Comprehensive ablations confirm the efficacy of each proposed component. All code models and datasets will be publicly available.
****
MRFS: Mutually Reinforcing Image Fusion and Segmentation
Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, Jiayi Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26974-26983
This paper proposes a coupled learning framework to break the performance bottleneck of infrared-visible image fusion and segmentation called MRFS. By leveraging the intrinsic consistency between vision and semantics it emphasizes mutual reinforcement rather than treating these tasks as separate issues. First we embed weakened information recovery and salient information integration into the image fusion task employing the CNN-based interactive gated mixed attention (IGM-Att) module to extract high-quality visual features. This aims to satisfy human visual perception producing fused images with rich textures high contrast and vivid colors. Second a transformer-based progressive cycle attention (PC-Att) module is developed to enhance semantic segmentation. It establishes single-modal self-reinforcement and cross-modal mutual complementarity enabling more accurate decisions in machine semantic perception. Then the cascade of IGM-Att and PC-Att couples image fusion and semantic segmentation tasks implicitly bringing vision-related and semantics-related features into closer alignment. Therefore they mutually provide learning priors to each other resulting in visually satisfying fused images and more accurate segmentation decisions. Extensive experiments on public datasets showcase the advantages of our method in terms of visual satisfaction and decision accuracy. The code is publicly available at https://github.com/HaoZhang1018/MRFS.
****
MemoNav: Working Memory Model for Visual Navigation
Hongxin Li, Zeyu Wang, Xu Yang, Yuran Yang, Shuqi Mei, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17913-17922
Image-goal navigation is a challenging task that requires an agent to navigate to a goal indicated by an image in unfamiliar environments. Existing methods utilizing diverse scene memories suffer from inefficient exploration since they use all historical observations for decision-making without considering the goal-relevant fraction. To address this limitation we present MemoNav a novel memory model for image-goal navigation which utilizes a working memory-inspired pipeline t

o improve navigation performance. Specifically we employ three types of navigati
on memory. The node features on a map are stored in the short-term memory (STM)
as these features are dynamically updated. A forgetting module then retains the
informative STM fraction to increase efficiency. We also introduce long-term mem
ory (LTM) to learn global scene representations by progressively aggregating STM
 features. Subsequently a graph attention module encodes the retained STM and th
e LTM to generate working memory (WM) which contains the scene features essentia
l for efficient navigation. The synergy among these three memory types boosts na
vigation performance by enabling the agent to learn and leverage goal-relevant s
cene features within a topological map. Our evaluation on multi-goal tasks demon
strates that MemoNav significantly outperforms previous methods across all diffi
culty levels in both Gibson and Matterport3D scenes. Qualitative results further
 illustrate that MemoNav plans more efficient routes.
****
Robust Depth Enhancement via Polarization Prompt Fusion Tuning
Kei Ikemura, Yiming Huang, Felix Heide, Zhaoxiang Zhang, Qifeng Chen, Chenyang L
ei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2024, pp. 20710-20720
Existing depth sensors are imperfect and may provide inaccurate depth values in
challenging scenarios such as in the presence of transparent or reflective objec
ts. In this work we present a general framework that leverages polarization imag
ing to improve inaccurate depth measurements from various depth sensors. Previou
s polarization-based depth enhancement methods focus on utilizing pure physics-b
ased formulas for a single sensor. In contrast our method first adopts a learnin
g-based strategy where a neural network is trained to estimate a dense and compl
ete depth map from polarization data and a sensor depth map from different senso
rs. To further improve the performance we propose a Polarization Prompt Fusion T
uning (PPFT) strategy to effectively utilize RGB-based models pre-trained on lar
ge-scale datasets as the size of the polarization dataset is limited to train a
strong model from scratch. We conducted extensive experiments on a public datase
t and the results demonstrate that the proposed method performs favorably compar
ed to existing depth enhancement baselines. Code and demos are available at http
s://lastbasket.github.io/PPFT/.
****
AssistGUI: Task-Oriented PC Graphical User Interface Automation
Difei Gao, Lei Ji, Zechen Bai, Mingyu Ouyang, Peiran Li, Dongxing Mao, Qinchen W
u, Weichen Zhang, Peiyi Wang, Xiangwu Guo, Hengxu Wang, Luowei Zhou, Mike Zheng
Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2024, pp. 13289-13298
Graphical User Interface (GUI) automation holds significant promise for assistin
g users with complex tasks thereby boosting human productivity. Existing works l
everaging Large Language Model (LLM) or LLM-based AI agents have shown capabilit
ies in automating tasks on Android and Web platforms. However these tasks are pr
imarily aimed at simple device usage and entertainment operations. This paper pr
esents a novel benchmark AssistGUI to evaluate whether models are capable of man
ipulating the mouse and keyboard on the Windows platform in response to user-req
uested tasks. We carefully collected a set of 100 tasks from nine widely-used so
ftware applications such as After Effects and MS Word each accompanied by the ne
cessary project files for better evaluation. Moreover we propose a multi-agent c
ollaboration framework which incorporates four agents to perform task decomposit
ion GUI parsing action generation and reflection. Our experimental results revea
l that our multi-agent collaboration mechanism outshines existing methods in per
formance. Nevertheless the potential remains substantial with the best model att
aining only a 46% success rate on our benchmark. We conclude with a thorough ana
lysis of the current methods' limitations setting the stage for future breakthro
ughs in this domain.
****
Adaptive Multi-Modal Cross-Entropy Loss for Stereo Matching
Peng Xu, Zhiyu Xiang, Chengyu Qiao, Jingyun Fu, Tianyu Pu; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5

135-5144
Despite the great success of deep learning in stereo matching recovering accurate disparity maps is still challenging. Currently L1 and cross-entropy are the two most widely used losses for stereo network training. Compared with the former the latter usually performs better thanks to its probability modeling and direct supervision to the cost volume. However how to accurately model the stereo ground-truth for cross-entropy loss remains largely under-explored. Existing works simply assume that the ground-truth distributions are uni-modal which ignores the fact that most of the edge pixels can be multi-modal. In this paper a novel adaptive multi-modal cross-entropy loss (ADL) is proposed to guide the networks to learn different distribution patterns for each pixel. Moreover we optimize the disparity estimator to further alleviate the bleeding or misalignment artifacts in inference. Extensive experimental results show that our method is generic and can help classic stereo networks regain state-of-the-art performance. In particular GANet with our method ranks 1st on both the KITTI 2015 and 2012 benchmarks among the published methods. Meanwhile excellent synthetic-to-realistic generalization performance can be achieved by simply replacing the traditional loss with ours. Code is available at https://github.com/xxxupeng/ADL.
****
Unlocking the Potential of Prompt-Tuning in Bridging Generalized and Personalized Federated Learning
Wenlong Deng, Christos Thrampoulidis, Xiaoxiao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6087-6097
Vision Transformers (ViT) and Visual Prompt Tuning (VPT) achieve state-of-the-art performance with improved efficiency in various computer vision tasks. This suggests a promising paradigm shift of adapting pre-trained ViT models to Federated Learning (FL) settings. However the challenge of data heterogeneity among FL clients presents a significant hurdle in effectively deploying ViT models. Existing Generalized FL (GFL) and Personalized FL (PFL) methods have limitations in balancing performance across both global and local data distributions. In this paper we present a novel algorithm SGPT that integrates GFL and PFL approaches by employing a unique combination of both shared and group-specific prompts. This design enables SGPT to capture both common and group-specific features. A key feature of SGPT is its prompt selection module which facilitates the training of a single global model capable of automatically adapting to diverse local client data distributions without the need for local fine-tuning. To effectively train the prompts we utilize block coordinate descent (BCD) learning from common feature information (shared prompts) and then more specialized knowledge (group prompts) iteratively. Theoretically we justify that learning the proposed prompts can reduce the gap between global and local performance. Empirically we conduct experiments on both label and feature heterogeneity settings in comparison with state-of-the-art baselines along with extensive ablation studies to substantiate the superior performance of SGPT.
****
Compact 3D Gaussian Representation for Radiance Field
Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, Eunbyung Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21719-21728
Neural Radiance Fields (NeRFs) have demonstrated remarkable potential in capturing complex 3D scenes with high fidelity. However one persistent challenge that hinders the widespread adoption of NeRFs is the computational bottleneck due to the volumetric rendering. On the other hand 3D Gaussian splatting (3DGS) has recently emerged as an alternative representation that leverages a 3D Gaussisan-based representation and adopts the rasterization pipeline to render the images rather than volumetric rendering achieving very fast rendering speed and promising image quality. However a significant drawback arises as 3DGS entails a substantial number of 3D Gaussians to maintain the high fidelity of the rendered images which requires a large amount of memory and storage. To address this critical issue we place a specific emphasis on two key objectives: reducing the number of Gaussian points without sacrificing performance and compressing the Gaussian attrib

utes such as view-dependent color and covariance. To this end we propose a learn able mask strategy that significantly reduces the number of Gaussians while pres erving high performance. In addition we propose a compact but effective represen tation of view-dependent color by employing a grid-based neural field rather tha n relying on spherical harmonics. Finally we learn codebooks to compactly repres ent the geometric attributes of Gaussian by vector quantization. With model comp ression techniques such as quantization and entropy coding we consistently show over 25x reduced storage and enhanced rendering speed while maintaining the qual ity of the scene representation compared to 3DGS. Our work provides a comprehens ive framework for 3D scene representation achieving high performance fast traini ng compactness and real-time rendering. Our project page is available at https:/ /maincold2.github.io/c3dgs/.
****
PaSCo: Urban 3D Panoptic Scene Completion with Uncertainty Awareness
Anh-Quan Cao, Angela Dai, Raoul de Charette; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14554-14564
We propose the task of Panoptic Scene Completion (PSC) which extends the recentl y popular Semantic Scene Completion (SSC) task with instance-level information t o produce a richer understanding of the 3D scene. Our PSC proposal utilizes a hy brid mask-based technique on the nonempty voxels from sparse multi-scale complet ions. Whereas the SSC literature overlooks uncertainty which is critical for rob otics applications we instead propose an efficient ensembling to estimate both v oxel-wise and instance-wise uncertainties along PSC. This is achieved by buildin g on a multi-input multi-output (MIMO) strategy while improving performance and yielding better uncertainty for little additional compute. Additionally we intro duce a technique to aggregate permutation-invariant mask predictions. Our experi ments demonstrate that our method surpasses all baselines in both Panoptic Scene Completion and uncertainty estimation on three large-scale autonomous driving d atasets. Our code and data are available at https://astra-vision.github.io/PaSCo .
****
GALA: Generating Animatable Layered Assets from a Single Scan
Taeksoo Kim, Byungjun Kim, Shunsuke Saito, Hanbyul Joo; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1535 -1545
We present GALA a framework that takes as input a single-layer clothed 3D human mesh and decomposes it into complete multi-layered 3D assets. The outputs can th en be combined with other assets to create novel clothed human avatars with any pose. Existing reconstruction approaches often treat clothed humans as a single- layer of geometry and overlook the inherent compositionality of humans with hair styles clothing and accessories thereby limiting the utility of the meshes for d own-stream applications. Decomposing a single-layer mesh into separate layers is a challenging task because it requires the synthesis of plausible geometry and texture for the severely occluded regions. Moreover even with successful decompo sition meshes are not normalized in terms of poses and body shapes failing coher ent composition with novel identities and poses. To address these challenges we propose to leverage the general knowledge of a pretrained 2D diffusion model as geometry and appearance prior for humans and other assets. We first separate the input mesh using the 3D surface segmentation extracted from multi-view 2D segme ntations. Then we synthesize the missing geometry of different layers in both po sed and canonical spaces using a novel pose-guided Score Distillation Sampling (SDS) loss. Once we complete inpainting high-fidelity 3D geometry we also apply t he same SDS loss to its texture to obtain the complete appearance including the initially occluded regions. Through a series of decomposition steps we obtain mu ltiple layers of 3D assets in a shared canonical space normalized in terms of po ses and human shapes hence supporting effortless composition to novel identities and reanimation with novel poses. Our experiments demonstrate the effectiveness of our approach for decomposition canonicalization and composition tasks compar ed to existing solutions.
****

LeGO: Leveraging a Surface Deformation Network for Animatable Stylized Face Gene
ration with One Example
Soyeon Yoon, Kwan Yun, Kwanggyoon Seo, Sihun Cha, Jung Eun Yoo, Junyong Noh; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2024, pp. 4505-4514
Recent advances in 3D face stylization have made significant strides in few to z
ero-shot settings. However the degree of stylization achieved by existing method
s is often not sufficient for practical applications because they are mostly bas
ed on statistical 3D Morphable Models (3DMM) with limited variations. To this en
d we propose a method that can produce a highly stylized 3D face model with desi
red topology. Our methods train a surface deformation network with 3DMM and tran
slate its domain to the target style using a paired exemplar. The network achiev
es stylization of the 3D face mesh by mimicking the style of the target using a
differentiable renderer and directional CLIP losses. Additionally during the inf
erence process we utilize a Mesh Agnostic Encoder (MAGE) that takes deformation
target a mesh of diverse topologies as input to the stylization process and enco
des its shape into our latent space. The resulting stylized face model can be an
imated by commonly used 3DMM blend shapes. A set of quantitative and qualitative
 evaluations demonstrate that our method can produce highly stylized face meshes
 according to a given style and output them in a desired topology. We also demon
strate example applications of our method including image-based stylized avatar
generation linear interpolation of geometric styles and facial animation of styl
ized avatars.
****
Frequency-Adaptive Dilated Convolution for Semantic Segmentation
Linwei Chen, Lin Gu, Dezhi Zheng, Ying Fu; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3414-3425
Dilated convolution which expands the receptive field by inserting gaps between
its consecutive elements is widely employed in computer vision. In this study we
 propose three strategies to improve individual phases of dilated convolution fr
om the view of spectrum analysis. Departing from the conventional practice of fi
xing a global dilation rate as a hyperparameter we introduce Frequency-Adaptive
Dilated Convolution (FADC) which dynamically adjusts dilation rates spatially ba
sed on local frequency components. Subsequently we design two plug-in modules to
 directly enhance effective bandwidth and receptive field size. The Adaptive Ker
nel (AdaKern) module decomposes convolution weights into low-frequency and high-
frequency components dynamically adjusting the ratio between these components on
 a per-channel basis. By increasing the high-frequency part of convolution weigh
ts AdaKern captures more high-frequency components thereby improving effective b
andwidth. The Frequency Selection (FreqSelect) module optimally balances high- a
nd low-frequency components in feature representations through spatially variant
 reweighting. It suppresses high frequencies in the background to encourage FADC
 to learn a larger dilation thereby increasing the receptive field for an expand
ed scope. Extensive experiments on segmentation and object detection consistentl
y validate the efficacy of our approach. The code is made publicly available at
https://github.com/Linwei-Chen/FADC.
****
3D Building Reconstruction from Monocular Remote Sensing Images with Multi-level
 Supervisions
Weijia Li, Haote Yang, Zhenghao Hu, Juepeng Zheng, Gui-Song Xia, Conghui He; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2024, pp. 27728-27737
3D building reconstruction from monocular remote sensing images is an important
and challenging research problem that has received increasing attention in recen
t years owing to its low cost of data acquisition and availability for large-sca
le applications. However existing methods rely on expensive 3D-annotated samples
 for fully-supervised training restricting their application to large-scale cros
s-city scenarios. In this work we propose MLS-BRN a multi-level supervised build
ing reconstruction network that can flexibly utilize training samples with diffe
rent annotation levels to achieve better reconstruction results in an end-to-end

manner. To alleviate the demand on full 3D supervision we design two new module
s Pseudo Building Bbox Calculator and Roof-Offset guided Footprint Extractor as
well as new tasks and training strategies for different types of samples. Experi
mental results on several public and new datasets demonstrate that our proposed
MLS-BRN achieves competitive performance using much fewer 3D-annotated samples a
nd significantly improves the footprint extraction and 3D reconstruction perform
ance compared with current state-of-the-art. The code and datasets of this work
will be released at https://github.com/opendatalab/MLS-BRN.git.
****
PhyScene: Physically Interactable 3D Scene Synthesis for Embodied AI
Yandan Yang, Baoxiong Jia, Peiyuan Zhi, Siyuan Huang; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16262-
16272
With recent developments in Embodied Artificial Intelligence (EAI) research ther
e has been a growing demand for high-quality large-scale interactive scene gener
ation. While prior methods in scene synthesis have prioritized the naturalness a
nd realism of the generated scenes the physical plausibility and interactivity o
f scenes have been largely left unexplored. To address this disparity we introdu
ce PhyScene a novel method dedicated to generating interactive 3D scenes charact
erized by realistic layouts articulated objects and rich physical interactivity
tailored for embodied agents. Based on a conditional diffusion model for capturi
ng scene layouts we devise novel physics- and interactivity-based guidance mecha
nisms that integrate constraints from object collision room layout and object re
achability. Through extensive experiments we demonstrate that PhyScene effective
ly leverages these guidance functions for physically interactable scene synthesi
s outperforming existing state-of-the-art scene synthesis methods by a large mar
gin. Our findings suggest that the scenes generated by PhyScene hold considerabl
e potential for facilitating diverse skill acquisition among agents within inter
active environments thereby catalyzing further advancements in embodied AI resea
rch.
****
Generative Latent Coding for Ultra-Low Bitrate Image Compression
Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, Yan Lu; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26088-
26098
Most existing image compression approaches perform transform coding in the pixel
space to reduce its spatial redundancy. However they encounter difficulties in
achieving both high-realism and high-fidelity at low bitrate as the pixel-space
distortion may not align with human perception. To address this issue we introdu
ce a Generative Latent Coding (GLC) architecture which performs transform coding
in the latent space of a generative vector-quantized variational auto-encoder (
VQ-VAE) instead of in the pixel space. The generative latent space is characteri
zed by greater sparsity richer semantic and better alignment with human percepti
on rendering it advantageous for achieving high-realism and high-fidelity compre
ssion. Additionally we introduce a categorical hyper module to reduce the bit co
st of hyper-information and a code-prediction-based supervision to enhance the s
emantic consistency. Experiments demonstrate that our GLC maintains high visual
quality with less than 0.04 bpp on natural images and less than 0.01 bpp on faci
al images. On the CLIC2020 test set we achieve the same FID as MS-ILLM with 45%
fewer bits. Furthermore the powerful generative latent space enables various app
lications built on our GLC pipeline such as image restoration and style transfer
.
****
Multiple View Geometry Transformers for 3D Human Pose Estimation
Ziwei Liao, Jialiang Zhu, Chunyu Wang, Han Hu, Steven L. Waslander; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
24, pp. 708-717
In this work we aim to improve the 3D reasoning ability of Transformers in multi
-view 3D human pose estimation. Recent works have focused on end-to-end learning
-based transformer designs which struggle to resolve geometric information accur

ately particularly during occlusion. Instead we propose a novel hybrid model MVG Former which has a series of geometric and appearance modules organized in an it erative manner. The geometry modules are learning-free and handle all viewpoint-dependent 3D tasks geometrically which notably improves the model's generalizati on ability. The appearance modules are learnable and are dedicated to estimating 2D poses from image signals end-to-end which enables them to achieve accurate e stimates even when occlusion occurs leading to a model that is both accurate and generalizable to new cameras and geometries. We evaluate our approach for both in-domain and out-of-domain settings where our model consistently outperforms st ate-of-the-art methods and especially does so by a significant margin in the out -of-domain setting. We will release the code and models: https://github.com/Xuns hanMan/MVGFormer.
****
SiTH: Single-view Textured Human Reconstruction with Image-Conditioned Diffusion
Hsuan- I Ho, Jie Song, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 538-549
A long-standing goal of 3D human reconstruction is to create lifelike and fully detailed 3D humans from single-view images. The main challenge lies in inferring unknown body shapes appearances and clothing details in areas not visible in th e images. To address this we propose SiTH a novel pipeline that uniquely integra tes an image-conditioned diffusion model into a 3D mesh reconstruction workflow. At the core of our method lies the decomposition of the challenging single-view reconstruction problem into generative hallucination and reconstruction subprob lems. For the former we employ a powerful generative diffusion model to hallucin ate unseen back-view appearance based on the input images. For the latter we lev erage skinned body meshes as guidance to recover full-body texture meshes from t he input and back-view images. SiTH requires as few as 500 3D human scans for tr aining while maintaining its generality and robustness to diverse images. Extens ive evaluations on two 3D human benchmarks including our newly created one highl ighted our method's superior accuracy and perceptual quality in 3D textured huma n reconstruction. Our code and evaluation benchmark is available at https://ait. ethz.ch/sith.
****
Distributionally Generative Augmentation for Fair Facial Attribute Classificatio n
Fengda Zhang, Qianpei He, Kun Kuang, Jiashuo Liu, Long Chen, Chao Wu, Jun Xiao, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2024, pp. 22797-22808
Facial Attribute Classification (FAC) holds substantial promise in widespread ap plications. However FAC models trained by traditional methodologies can be unfai r by exhibiting accuracy inconsistencies across varied data subpopulations. This unfairness is largely attributed to bias in data where some spurious attributes (e.g. Male) statistically correlate with the target attribute (e.g. Smiling). M ost of existing fairness-aware methods rely on the labels of spurious attributes which may be unavailable in practice. This work proposes a novel generation-bas ed two-stage framework to train a fair FAC model on biased data without addition al annotation. Initially we identify the potential spurious attributes based on generative models. Notably it enhances interpretability by explicitly showing th e spurious attributes in image space. Following this for each image we first edi t the spurious attributes with a random degree sampled from a uniform distributi on while keeping target attribute unchanged. Then we train a fair FAC model by f ostering model invariance to these augmentation. Extensive experiments on three common datasets demonstrate the effectiveness of our method in promoting fairnes s in FAC without compromising accuracy. Codes are in https://github.com/heqianpe i/DiGA.
****
DynVideo-E: Harnessing Dynamic NeRF for Large-Scale Motion- and View-Change Huma n-Centric Video Editing
Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Juss i Keppo, Ying Shan, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on C

omputer Vision and Pattern Recognition (CVPR), 2024, pp. 7664-7674
Despite recent progress in diffusion-based video editing existing methods are limited to short-length videos due to the contradiction between long-range consistency and frame-wise editing. Prior attempts to address this challenge by introducing video-2D representations encounter significant difficulties with large motion- and view-change videos especially in human-centric scenarios. To overcome this we propose to introduce the dynamic Neural Radiance Fields (NeRF) as the innovative video representation where the editing can be performed in the 3D spaces and propagated to the entire video via the deformation field. To provide consistent and controllable editing we propose the image-based video-NeRF editing pipeline with a set of innovative designs including multi-view multi-pose Score Distillation Sampling (SDS) from both the 2D personalized diffusion prior and 3D diffusion prior reconstruction losses text-guided local parts super-resolution and style transfer. Extensive experiments demonstrate that our method dubbed as DynVideo-E significantly outperforms SOTA approaches on two challenging datasets by a large margin of 50%   95% for human preference. Code will be released at https://showlab.github.io/DynVideo-E/.
****
Real-Time Neural BRDF with Spherically Distributed Primitives
Yishun Dou, Zhong Zheng, Qiaoqiao Jin, Bingbing Ni, Yugang Chen, Junxiang Ke; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4337-4346
We propose a neural reflectance model (NeuBRDF) that offers highly versatile material representation yet with light memory and neural computation consumption towards achieving real-time rendering. The results depicted in Fig. 1 rendered at full HD resolution on a contemporary desktop machine demonstrate that our system achieves real-time performance with a wide variety of appearances which is approached by the following two designs. Firstly recognizing that the bidirectional reflectance is distributed in a sparse high-dimensional space we propose to project the BRDF into two low-dimensional components i.e. two hemisphere feature-grids for incoming and outgoing directions respectively. Secondly we distribute learnable neural reflectance primitives on our highly-tailored spherical surface grid. These primitives offer informative features for each hemisphere component and reduce the complexity of the feature learning network leading to fast evaluation. These primitives are centrally stored in a codebook and can be shared across multiple grids and even across materials based on low-cost indices stored in material-specific spherical surface grids. Our NeuBRDF agnostic to the material provides a unified framework for representing a variety of materials consistently. Comprehensive experimental results on measured BRDF compression Monte Carlo simulated BRDF acceleration and extension to spatially varying effects demonstrate the superior quality and generalizability achieved by the proposed scheme.
****
Harnessing Meta-Learning for Improving Full-Frame Video Stabilization
Muhammad Kashif Ali, Eun Woo Im, Dongjin Kim, Tae Hyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12605-12614
Video stabilization is a longstanding computer vision problem particularly pixel-level synthesis solutions for video stabilization which synthesize full frames add to the complexity of this task. These techniques aim to stabilize videos by synthesizing full frames while enhancing the stability of the considered video. This intensifies the complexity of the task due to the distinct mix of unique motion profiles and visual content present in each video sequence making robust generalization with fixed parameters difficult. In our study we introduce a novel approach to enhance the performance of pixel-level synthesis solutions for video stabilization by adapting these models to individual input video sequences. The proposed adaptation exploits low-level visual cues accessible during test-time to improve both the stability and quality of resulting videos. We highlight the efficacy of our methodology of "test-time adaptation" through simple fine-tuning of one of these models followed by significant stability gain via the integration of meta-learning techniques. Notably significant improvement is achieved with

only a single adaptation step. The versatility of the proposed algorithm is demonstrated by consistently improving the performance of various pixel-level synthesis models for video stabilization in real-world scenarios.
****
VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models
Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7310-7320
Text-to-video generation aims to produce a video based on a given prompt. Recently several commercial video models have been able to generate plausible videos with minimal noise excellent details and high aesthetic scores. However these models rely on large-scale well-filtered high-quality videos that are not accessible to the community. Many existing research works which train models using the low-quality WebVid-10M dataset struggle to generate high-quality videos because the models are optimized to fit WebVid-10M. In this work we explore the training scheme of video models extended from Stable Diffusion and investigate the feasibility of leveraging low-quality videos and synthesized high-quality images to obtain a high-quality video model. We first analyze the connection between the spatial and temporal modules of video models and the distribution shift to low-quality videos. We observe that full training of all modules results in a stronger coupling between spatial and temporal modules than only training temporal modules. Based on this stronger coupling we shift the distribution to higher quality without motion degradation by finetuning spatial modules with high-quality images resulting in a generic high-quality video model. Evaluations are conducted to demonstrate the superiority of the proposed method particularly in picture quality motion and concept composition.
****
From SAM to CAMs: Exploring Segment Anything Model for Weakly Supervised Semantic Segmentation
Hyeokjun Kweon, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19499-19509
Weakly Supervised Semantic Segmentation (WSSS) aims to learn the concept of segmentation using image-level class labels. Recent WSSS works have shown promising results by using the Segment Anything Model (SAM) a foundation model for segmentation during the inference phase. However we observe that these methods can still be vulnerable to the noise of class activation maps (CAMs) serving as initial seeds. As a remedy this paper introduces From-SAM-to-CAMs (S2C) a novel WSSS framework that directly transfers the knowledge of SAM to the classifier during the training process enhancing the quality of CAMs itself. S2C comprises SAM-segment Contrasting (SSC) and a CAM-based prompting module (CPM) which exploit SAM at the feature and logit levels respectively. SSC performs prototype-based contrasting using SAM's automatic segmentation results. It constrains each feature to be close to the prototype of its segment and distant from prototypes of the others. Meanwhile CPM extracts prompts from the CAM of each class and uses them to generate class-specific segmentation masks through SAM. The masks are aggregated into unified self-supervision based on the confidence score designed to consider the reliability of both SAM and CAMs. S2C achieves a new state-of-the-art performance across all benchmarks outperforming existing studies by significant margins. The code is available at https://github.com/sangrockEG/S2C.
****
Boosting Flow-based Generative Super-Resolution Models via Learned Prior
Li-Yuan Tsao, Yi-Chen Lo, Chia-Che Chang, Hao-Wei Chen, Roy Tseng, Chien Feng, Chun-Yi Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26005-26015
Flow-based super-resolution (SR) models have demonstrated astonishing capabilities in generating high-quality images. However these methods encounter several challenges during image generation such as grid artifacts exploding inverses and suboptimal results due to a fixed sampling temperature. To overcome these issues this work introduces a conditional learned prior to the inference phase of a flo

w-based SR model. This prior is a latent code predicted by our proposed latent module conditioned on the low-resolution image which is then transformed by the flow model into an SR image. Our framework is designed to seamlessly integrate with any contemporary flow-based SR model without modifying its architecture or pre-trained weights. We evaluate the effectiveness of our proposed framework through extensive experiments and ablation analyses. The proposed framework successfully addresses all the inherent issues in flow-based SR models and enhances their performance in various SR scenarios. Our code is available at: https://github.com/liyuantsao/FlowSR-LP
****
How to Handle Sketch-Abstraction in Sketch-Based Image Retrieval?
Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16859-16869
In this paper we propose a novel abstraction-aware sketch-based image retrieval framework capable of handling sketch abstraction at varied levels. Prior works had mainly focused on tackling sub-factors such as drawing style and order we instead attempt to model abstraction as a whole and propose feature-level and retrieval granularity-level designs so that the system builds into its DNA the necessary means to interpret abstraction. On learning abstraction-aware features we for the first-time harness the rich semantic embedding of pre-trained StyleGAN model together with a novel abstraction-level mapper that deciphers the level of abstraction and dynamically selects appropriate dimensions in the feature matrix correspondingly to construct a feature matrix embedding that can be freely traversed to accommodate different levels of abstraction. For granularity-level abstraction understanding we dictate that the retrieval model should not treat all abstraction-levels equally and introduce a differentiable surrogate Acc.@q loss to inject that understanding into the system. Different to the gold-standard triplet loss our Acc.@q loss uniquely allows a sketch to narrow/broaden its focus in terms of how stringent the evaluation should be - the more abstract a sketch the less stringent (higher q). Extensive experiments depict our method to outperform existing state-of-the-arts in standard SBIR tasks along with challenging scenarios like early retrieval forensic sketch-photo matching and style-invariant retrieval.
****
What You See is What You GAN: Rendering Every Pixel for High-Fidelity Geometry in 3D GANs
Alex Trevithick, Matthew Chan, Towaki Takikawa, Umar Iqbal, Shalini De Mello, Manmohan Chandraker, Ravi Ramamoorthi, Koki Nagano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22765-22775
3D-aware Generative Adversarial Networks (GANs) have shown remarkable progress in learning to generate multi-view-consistent images and 3D geometries of scenes from collections of 2D images via neural volume rendering. Yet the significant memory and computational costs of dense sampling in volume rendering have forced 3D GANs to adopt patch-based training or employ low-resolution rendering with post-processing 2D super resolution which sacrifices multiview consistency and the quality of resolved geometry. Consequently 3D GANs have not yet been able to fully resolve the rich 3D geometry present in 2D images. In this work we propose techniques to scale neural volume rendering to the much higher resolution of native 2D images thereby resolving fine-grained 3D geometry with unprecedented detail. Our approach employs learning-based samplers for accelerating neural rendering for 3D GAN training using up to 5 times fewer depth samples. This enables us to explicitly "render every pixel" of the full-resolution image during training and inference without post-processing superresolution in 2D. Together with our strategy to learn high-quality surface geometry our method synthesizes high-resolution 3D geometry and strictly view-consistent images while maintaining image quality on par with baselines relying on post-processing super resolution. We demonstrate state-of-the-art 3D gemetric quality on FFHQ and AFHQ setting a new standard for unsupervised learning of 3D shapes in 3D GANs.

****
Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer
Jiwoo Chung, Sangeek Hyun, Jae-Pil Heo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8795-8805
Despite the impressive generative capabilities of diffusion models existing diffusion model-based style transfer methods require inference-stage optimization (e.g. fine-tuning or textual inversion of style) which is time-consuming or fails to leverage the generative ability of large-scale diffusion models. To address these issues we introduce a novel artistic style transfer method based on a pre-trained large-scale diffusion model without any optimization. Specifically we manipulate the features of self-attention layers as the way the cross-attention mechanism works; in the generation process substituting the key and value of content with those of style image. This approach provides several desirable characteristics for style transfer including 1) preservation of content by transferring similar styles into similar image patches and 2) transfer of style based on similarity of local texture (e.g. edge) between content and style images. Furthermore we introduce query preservation and attention temperature scaling to mitigate the issue of disruption of original content and initial latent Adaptive Instance Normalization (AdaIN) to deal with the disharmonious color (failure to transfer the colors of style). Our experimental results demonstrate that our proposed method surpasses state-of-the-art methods in both conventional and diffusion-based style transfer baselines.
****
Towards Robust Learning to Optimize with Theoretical Guarantees
Qingyu Song, Wei Lin, Juncheng Wang, Hong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27498-27506
Learning to optimize (L2O) is an emerging technique to solve mathematical optimization problems with learning-based methods. Although with great success in many real-world scenarios such as wireless communications computer networks and electronic design existing L2O works lack theoretical demonstration of their performance and robustness in out-of-distribution (OOD) scenarios. We address this gap by providing comprehensive proofs. First we prove a sufficient condition for a robust L2O model with homogeneous convergence rates over all In-Distribution (InD) instances. We assume an L2O model achieves robustness for an InD scenario. Based on our proposed methodology of aligning OOD problems to InD problems we also demonstrate that the L2O model's convergence rate in OOD scenarios will deteriorate by an equation of the L2O model's input features. Moreover we propose an L2O model with a concise gradient-only feature construction and a novel gradient-based history modeling method. Numerical simulation demonstrates that our proposed model outperforms the state-of-the-art baseline in both InD and OOD scenarios and achieves up to 10 xconvergence speedup. The code of our method can be found from https://github.com/NetX-lab/GoMathL2O-Official.
****
Differentiable Neural Surface Refinement for Modeling Transparent Objects
Weijian Deng, Dylan Campbell, Chunyi Sun, Shubham Kanitkar, Matthew E. Shaffer, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20268-20277
Neural implicit surface reconstruction leveraging volume rendering has led to significant advances in multi-view reconstruction. However results for transparent objects can be very poor primarily because the rendering function fails to account for the intricate light transport induced by refraction and reflection. In this study we introduce transparent neural surface refinement (TNSR) a novel surface reconstruction framework that explicitly incorporates physical refraction and reflection tracing. Beginning with an initial approximate surface our method employs sphere tracing combined with Snell's law to cast both reflected and refracted rays. Central to our proposal is an innovative differentiable technique devised to allow signals from the photometric evidence to propagate back to the surface model by considering how the surface bends and reflects light rays. This allows us to connect surface refinement with volume rendering enabling end-to-end

optimization solely on multi-view RGB images. In our experiments TNSR demonstrates significant improvements in novel view synthesis and geometry estimation of transparent objects without prior knowledge of the refractive index.
****
OrthCaps: An Orthogonal CapsNet with Sparse Attention Routing and Pruning
Xinyu Geng, Jiaming Wang, Jiawei Gong, Yuerong Xue, Jun Xu, Fanglin Chen, Xiaolin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6037-6046
Redundancy is a persistent challenge in Capsule Networks (CapsNet) leading to high computational costs and parameter counts. Although previous studies have introduced pruning after the initial capsule layer dynamic routing's fully connected nature and non-orthogonal weight matrices reintroduce redundancy in deeper layers. Besides dynamic routing requires iterating to converge further increasing computational demands. In this paper we propose an Orthogonal Capsule Network (OrthCaps) to reduce redundancy improve routing performance and decrease parameter counts. Firstly an efficient pruned capsule layer is introduced to discard redundant capsules. Secondly dynamic routing is replaced with orthogonal sparse attention routing eliminating the need for iterations and fully connected structures. Lastly weight matrices during routing are orthogonalized to sustain low capsule similarity which is the first approach to use Householder orthogonal decomposition to enforce orthogonality in CapsNet. Our experiments on baseline datasets affirm the efficiency and robustness of OrthCaps in classification tasks in which ablation studies validate the criticality of each component. OrthCaps-Shallow outperforms other Capsule Network benchmarks on four datasets utilizing only 110k parameters - a mere 1.25% of a standard Capsule Network's total. To the best of our knowledge it achieves the smallest parameter count among existing Capsule Networks. Similarly OrthCaps-Deep demonstrates competitive performance across four datasets utilizing only 1.2% of the parameters required by its counterparts.
****
ProS: Prompting-to-simulate Generalized knowledge for Universal Cross-Domain Retrieval
Kaipeng Fang, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Zhi-Qi Cheng, Xiyao Li, Heng Tao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17292-17301
The goal of Universal Cross-Domain Retrieval (UCDR) is to achieve robust performance in generalized test scenarios wherein data may belong to strictly unknown domains and categories during training. Recently pre-trained models with prompt tuning have shown strong generalization capabilities and attained noteworthy achievements in various downstream tasks such as few-shot learning and video-text retrieval. However applying them directly to UCDR may not be sufficient to handle both domain shift (i.e. adapting to unfamiliar domains) and semantic shift (i.e. transferring to unknown categories). To this end we propose Prompting-to-Simulate (ProS) the first method to apply prompt tuning for UCDR. ProS employs a two-step process to simulate Content-aware Dynamic Prompts (CaDP) which can impact models to produce generalized features for UCDR. Concretely in Prompt Units Learning stage we introduce two Prompt Units to individually capture domain and semantic knowledge in a mask-and-align way. Then in Context-aware Simulator Learning stage we train a Content-aware Prompt Simulator under a simulated test scenario to produce the corresponding CaDP. Extensive experiments conducted on three benchmark datasets show that our method achieves new state-of-the-art performance without bringing excessive parameters. Code is available at https://github.com/fangkaipeng/ProS
****
Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks
Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, Lu Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4818-4829
We introduce Florence-2 a novel vision foundation model with a unified prompt-based representation for various computer vision and vision-language tasks. While existing large vision models excel in transfer learning they struggle to perform

diverse tasks with simple instructions a capability that implies handling the c omplexity of various spatial hierarchy and semantic granularity. Florence-2 was designed to take text-prompt as task instructions and generate desirable results in text forms whether it be captioning object detection grounding or segmentati on. This multi-task learning setup demands large-scale high-quality annotated da ta. To this end we co-developed FLD-5B that consists of 5.4 billion comprehensiv e visual annotations on 126 million images using an iterative strategy of automa ted image annotation and model refinement. We adopted a sequence-to-sequence str ucture to train Florence-2 to perform versatile and comprehensive vision tasks. Extensive evaluations on numerous tasks demonstrated Florence-2 to be a strong v ision foundation model contender with unprecedented zero-shot and fine-tuning ca pabilities.
****
NeRF On-the-go: Exploiting Uncertainty for Distractor-free NeRFs in the Wild
Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, Songyou Peng; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8931-8940
Neural Radiance Fields (NeRFs) have shown remarkable success in synthesizing pho torealistic views from multi-view images of static scenes but face challenges in dynamic real-world environments with distractors like moving objects shadows an d lighting changes. Existing methods manage controlled environments and low occl usion ratios but fall short in render quality especially under high occlusion sc enarios. In this paper we introduce NeRF On-the-go a simple yet effective approa ch that enables the robust synthesis of novel views in complex in-the-wild scene s from only casually captured image sequences. Delving into uncertainty our meth od not only efficiently eliminates distractors even when they are predominant in captures but also achieves a notably faster convergence speed. Through comprehe nsive experiments on various scenes our method demonstrates a significant improv ement over state-of-the-art techniques. This advancement opens new avenues for N eRF in diverse and dynamic real-world applications.
****
3D Human Pose Perception from Egocentric Stereo Videos
Hiroyasu Akada, Jian Wang, Vladislav Golyanik, Christian Theobalt; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 767-776
While head-mounted devices are becoming more compact they provide egocentric vie ws with significant self-occlusions of the device user. Hence existing methods o ften fail to accurately estimate complex 3D poses from egocentric views. In this work we propose a new transformer-based framework to improve egocentric stereo 3D human pose estimation which leverages the scene information and temporal cont ext of egocentric stereo videos. Specifically we utilize 1) depth features from our 3D scene reconstruction module with uniformly sampled windows of egocentric stereo frames and 2) human joint queries enhanced by temporal features of the vi deo inputs. Our method is able to accurately estimate human poses even in challe nging scenarios such as crouching and sitting. Furthermore we introduce two new benchmark datasets i.e. UnrealEgo2 and UnrealEgo-RW (RealWorld). UnrealEgo2 is a large-scale in-the-wild dataset captured in synthetic 3D scenes. UnrealEgo-RW i s a real-world dataset captured with our newly developed device. The proposed da tasets offer a much larger number of egocentric stereo views with a wider variet y of human motions than the existing datasets allowing comprehensive evaluation of existing and upcoming methods. Our extensive experiments show that the propos ed approach significantly outperforms previous methods. UnrealEgo2 UnrealEgo-RW and trained models are available on our project page and Benchmark Challenge.
****
Grid Diffusion Models for Text-to-Video Generation
Taegyeong Lee, Soyeong Kwon, Taehwan Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8734-8743
Recent advances in the diffusion models have significantly improved text-to-imag e generation. However generating videos from text is a more challenging task tha n generating images from text due to the much larger dataset and higher computat

ional cost required. Most existing video generation methods use either a 3D U-Net architecture that considers the temporal dimension or autoregressive generation. These methods require large datasets and are limited in terms of computational costs compared to text-to-image generation. To tackle these challenges we propose a simple but effective novel grid diffusion for text-to-video generation without temporal dimension in architecture and a large text-video paired dataset. We can generate a high-quality video using a fixed amount of GPU memory regardless of the number of frames by representing the video as a grid image. Additionally since our method reduces the dimensions of the video to the dimensions of the image various image-based methods can be applied to videos such as text-guided video manipulation from image manipulation. Our proposed method outperforms the existing methods in both quantitative and qualitative evaluations demonstrating the suitability of our model for real-world video generation.
****
Boosting Object Detection with Zero-Shot Day-Night Domain Adaptation
Zhipeng Du, Miaojing Shi, Jiankang Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12666-12676
Detecting objects in low-light scenarios presents a persistent challenge as detectors trained on well-lit data exhibit significant performance degradation on low-light data due to low visibility. Previous methods mitigate this issue by exploring image enhancement or object detection techniques with real low-light image datasets. However the progress is impeded by the inherent difficulties about collecting and annotating low-light images. To address this challenge we propose to boost low-light object detection with zero-shot day-night domain adaptation which aims to generalize a detector from well-lit scenarios to low-light ones without requiring real low-light data. Revisiting Retinex theory in the low-level vision we first design a reflectance representation learning module to learn Retinex-based illumination invariance in images with a carefully designed illumination invariance reinforcement strategy. Next an interchange-redecomposition-coherence procedure is introduced to improve over the vanilla Retinex image decomposition process by performing two sequential image decompositions and introducing a redecomposition cohering loss. Extensive experiments on ExDark DARK FACE and CODaN datasets show strong low-light generalizability of our method. Our code is available at https://github.com/ZPDu/DAI-Net.
****
LucidDreamer: Towards High-Fidelity Text-to-3D Generation via Interval Score Matching
Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, Yingcong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6517-6526
The recent advancements in text-to-3D generation mark a significant milestone in generative models unlocking new possibilities for creating imaginative 3D assets across various real-world scenarios. While recent advancements in text-to-3D generation have shown promise they often fall short in rendering detailed and high-quality 3D models. This problem is especially prevalent as many methods base themselves on Score Distillation Sampling (SDS). This paper identifies a notable deficiency in SDS that it brings inconsistent and low-quality updating direction for the 3D model causing the over-smoothing effect. To address this we propose a novel approach called Interval Score Matching (ISM). ISM employs deterministic diffusing trajectories and utilizes interval-based score matching to counteract over-smoothing. Furthermore we incorporate 3D Gaussian Splatting into our text-to-3D generation pipeline. Extensive experiments show that our model largely outperforms the state-of-the-art in quality and training efficiency.
****
PTM-VQA: Efficient Video Quality Assessment Leveraging Diverse PreTrained Models from the Wild
Kun Yuan, Hongbo Liu, Mading Li, Muyi Sun, Ming Sun, Jiachao Gong, Jinhua Hao, Chao Zhou, Yansong Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2835-2845
Video quality assessment (VQA) is a challenging problem due to the numerous fact

ors that can affect the perceptual quality of a video e.g. content attractivenes s distortion type motion pattern and level. However annotating the Mean opinion score (MOS) for videos is expensive and time-consuming which limits the scale of VQA datasets and poses a significant obstacle for deep learning-based methods. In this paper we propose a VQA method named PTM-VQA which leverages PreTrained M odels to transfer knowledge from models pretrained on various pre-tasks enabling benefits for VQA from different aspects. Specifically we extract features of vi deos from different pretrained models with frozen weights and integrate them to generate representation. Since these models possess various fields of knowledge and are often trained with labels irrelevant to quality we propose an Intra-Cons istency and Inter-Divisibility (ICID) loss to impose constraints on features ext racted by multiple pretrained models. The intra-consistency constraint ensures t hat features extracted by different pretrained models are in the same unified qu ality-aware latent space while the inter-divisibility introduces pseudo clusters based on the annotation of samples and tries to separate features of samples fr om different clusters. Furthermore with a constantly growing number of pretraine d models it is crucial to determine which models to use and how to use them. To address this problem we propose an efficient scheme to select suitable candidate s. Models with better clustering performance on VQA datasets are chosen to be ou r candidates. Extensive experiments demonstrate the effectiveness of the propose d method.
****
Versatile Medical Image Segmentation Learned from Multi-Source Datasets via Mode l Self-Disambiguation
Xiaoyang Chen, Hao Zheng, Yuemeng Li, Yuncong Ma, Liang Ma, Hongming Li, Yong Fa n; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2024, pp. 11747-11756
A versatile medical image segmentation model applicable to images acquired with diverse equipment and protocols can facilitate model deployment and maintenance. However building such a model typically demands a large diverse and fully annot ated dataset which is challenging to obtain due to the labor-intensive nature of data curation. To address this challenge we propose a cost-effective alternativ e that harnesses multi-source data with only partial or sparse segmentation labe ls for training substantially reducing the cost of developing a versatile model. We devise strategies for model self-disambiguation prior knowledge incorporatio n and imbalance mitigation to tackle challenges associated with inconsistently l abeled multi-source data including label ambiguity and modality dataset and clas s imbalances. Experimental results on a multi-modal dataset compiled from eight different sources for abdominal structure segmentation have demonstrated the eff ectiveness and superior performance of our method compared to state-of-the-art a lternative approaches. We anticipate that its cost-saving features which optimiz e the utilization of existing annotated data and reduce annotation efforts for n ew data will have a significant impact in the field.
****
Improving Generalization via Meta-Learning on Hard Samples
Nishant Jain, Arun S. Suggala, Pradeep Shenoy; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27600-27609
Learned reweighting (LRW) approaches to supervised learning use an optimization criterion to assign weights for training instances in order to maximize performa nce on a representative validation dataset. We pose and formalize the problem of optimized selection of the validation set used in LRW training to improve class ifier generalization. In particular we show that using hard-to-classify instance s in the validation set has both a theoretical connection to and strong empirica l evidence of generalization. We provide an efficient algorithm for training thi s meta-optimized model as well as a simple train-twice heuristic for careful com parative study. We demonstrate that LRW with easy validation data performs consi stently worse than LRW with hard validation data establishing the validity of ou r meta-optimization problem. Our proposed algorithm outperforms a wide range of baselines on a range of datasets and domain shift challenges (Imagenet-1K CIFAR-100 Clothing-1M CAMELYON WILDS etc.) with 1% gains using VIT-B on Imagenet. We

also show that using naturally hard examples for validation (Imagenet-R / Imagen et-A) in LRW training for Imagenet improves performance on both clean and natura lly hard test instances by 1-2%. Secondary analyses show that using hard validat ion data in an LRW framework improves margins on test data hinting at the mechan ism underlying our empirical gains. We believe this work opens up new research d irections for the meta-optimization of meta-learning in a supervised learning co ntext.
****
Align and Aggregate: Compositional Reasoning with Video Alignment and Answer Agg regation for Video Question-Answering
Zhaohe Liao, Jiangtong Li, Li Niu, Liqing Zhang; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13395-13404
Despite the recent progress made in Video Question-Answering (VideoQA) these met hods typically function as black-boxes making it difficult to understand their r easoning processes and perform consistent compositional reasoning. To address th ese challenges we propose a model-agnostic Video Alignment and Answer Aggregatio n (VA3) framework which is capable of enhancing both compositional consistency a nd accuracy of existing VidQA methods by integrating video aligner and answer ag gregator modules. The video aligner hierarchically selects the relevant video cl ips based on the question while the answer aggregator deduces the answer to the question based on its sub-questions with compositional consistency ensured by th e information flow along the question decompose graph and the contrastive learni ng strategy. We evaluate our framework on three settings of the AGQA-Decomp data set with three baseline methods and propose new metrics to measure the compositi onal consistency of VidQA methods more comprehensively. Moreover we propose a la rge language model (LLM) based automatic question decompose pipeline to apply ou r framework on any VidQA data. We extend MSVD and NExT-QA datasets with it to ev aluate such scheme and our VA3 framework on broader scenarios. Extensive experim ents show that our framework improves both compositional consistency and accurac y of existing methods leading to more interpretable models in real-world applica tions.
****
REACTO: Reconstructing Articulated Objects from a Single Video
Chaoyue Song, Jiacheng Wei, Chuan Sheng Foo, Guosheng Lin, Fayao Liu; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5384-5395
In this paper we address the challenge of reconstructing general articulated 3D objects from a single video. Existing works employing dynamic neural radiance fi elds have advanced the modeling of articulated objects like humans and animals f rom videos but face challenges with piece-wise rigid general articulated objects due to limitations in their deformation models. To tackle this we propose Quasi -Rigid Blend Skinning a novel deformation model that enhances the rigidity of ea ch part while maintaining flexible deformation of the joints. Our primary insigh t combines three distinct approaches: 1) an enhanced bone rigging system for imp roved component modeling 2) the use of quasi-sparse skinning weights to boost pa rt rigidity and reconstruction fidelity and 3) the application of geodesic point assignment for precise motion and seamless deformation. Our method outperforms previous works in producing higher-fidelity 3D reconstructions of general articu lated objects as demonstrated on both real and synthetic datasets. Project page: https://chaoyuesong.github.io/REACTO.
****
Egocentric Whole-Body Motion Capture with FisheyeViT and Diffusion-Based Motion Refinement
Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang , Thabo Beeler, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2024, pp. 777-787
In this work we explore egocentric whole-body motion capture using a single fish eye camera which simultaneously estimates human body and hand motion. This task presents significant challenges due to three factors: the lack of high-quality d atasets fisheye camera distortion and human body self-occlusion. To address thes

e challenges we propose a novel approach that leverages FisheyeViT to extract fisheye image features which are subsequently converted into pixel-aligned 3D heatmap representations for 3D human body pose prediction. For hand tracking we incorporate dedicated hand detection and hand pose estimation networks for regressing 3D hand poses. Finally we develop a diffusion-based whole-body motion prior model to refine the estimated whole-body motion while accounting for joint uncertainties. To train these networks we collect a large synthetic dataset EgoWholeBody comprising 840000 high-quality egocentric images captured across a diverse range of whole-body motion sequences. Quantitative and qualitative evaluations demonstrate the effectiveness of our method in producing high-quality whole-body motion estimates from a single egocentric camera.
****
Language Embedded 3D Gaussians for Open-Vocabulary Scene Understanding
Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, Shao-Hua Guan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5333-5343
Open-vocabulary querying in 3D space is challenging but essential for scene understanding tasks such as object localization and segmentation. Language-embedded scene representations have made progress by incorporating language features into 3D spaces. However their efficacy heavily depends on neural networks that are resource-intensive in training and rendering. Although recent 3D Gaussians offer efficient and high-quality novel view synthesis directly embedding language features in them leads to prohibitive memory usage and decreased performance. In this work we introduce Language Embedded 3D Gaussians a novel scene representation for open-vocabulary query tasks. Instead of embedding high-dimensional raw semantic features on 3D Gaussians we propose a dedicated quantization scheme that drastically alleviates the memory requirement and a novel embedding procedure that achieves smoother yet high accuracy query countering the multi-view feature inconsistencies and the high-frequency inductive bias in point-based representations. Our comprehensive experiments show that our representation achieves the best visual quality and language querying accuracy across current language-embedded representations while maintaining real-time rendering frame rates on a single desktop GPU.
****
Towards Automated Movie Trailer Generation
Dawit Mureja Argaw, Mattia Soldan, Alejandro Pardo, Chen Zhao, Fabian Caba Heilbron, Joon Son Chung, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7445-7454
Movie trailers are an essential tool for promoting films and attracting audiences. However the process of creating trailers can be time-consuming and expensive. To streamline this process we propose an automatic trailer generation framework that generates plausible trailers from a full movie by automating shot selection and composition. Our approach draws inspiration from machine translation techniques and models the movies and trailers as sequences of shots thus formulating the trailer generation problem as a sequence-to-sequence task. We introduce Trailer Generation Transformer (TGT) a deep-learning framework utilizing an encoder-decoder architecture. TGT movie encoder is tasked with contextualizing each movie shot representation via self-attention while the autoregressive trailer decoder predicts the feature representation of the next trailer shot accounting for the relevance of shots' temporal order in trailers. Our TGT significantly outperforms previous methods on a comprehensive suite of metrics.
****
Differentiable Information Bottleneck for Deterministic Multi-view Clustering
Xiaoqiang Yan, Zhixiang Jin, Fengshou Han, Yangdong Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27435-27444
In recent several years the information bottleneck (IB) principle provides an information-theoretic framework for deep multi-view clustering (MVC) by compressing multi-view observations while preserving the relevant information of multiple views. Although existing IB-based deep MVC methods have achieved huge success th

ey rely on variational approximation and distribution assumption to estimate the lower bound of mutual information which is a notoriously hard and impractical problem in high-dimensional multi-view spaces. In this work we propose a new differentiable information bottleneck (DIB) method which provides a deterministic and analytical MVC solution by fitting the mutual information without the necessity of variational approximation. Specifically we first propose to directly fit the mutual information of high-dimensional spaces by leveraging normalized kernel Gram matrix which does not require any auxiliary neural estimator to estimate the lower bound of mutual information. Then based on the new mutual information measurement a deterministic multi-view neural network with analytical gradients is explicitly trained to parameterize IB principle which derives a deterministic compression of input variables from different views. Finally a triplet consistency discovery mechanism is devised which is capable of mining the feature consistency cluster consistency and joint consistency based on the deterministic and compact representations. Extensive experimental results show the superiority of our DIB method on 6 benchmarks compared with 13 state-of-the-art baselines.
****
Sheared Backpropagation for Fine-tuning Foundation Models
Zhiyuan Yu, Li Shen, Liang Ding, Xinmei Tian, Yixin Chen, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5883-5892
Fine-tuning is the process of extending the training of pre-trained models on specific target tasks thereby significantly enhancing their performance across various applications. However fine-tuning often demands large memory consumption posing a challenge for low-memory devices that some previous memory-efficient fine-tuning methods attempted to mitigate by pruning activations for gradient computation albeit at the cost of significant computational overhead from the pruning processes during training. To address these challenges we introduce PreBackRazor a novel activation pruning scheme offering both computational and memory efficiency through a sparsified backpropagation strategy which judiciously avoids unnecessary activation pruning and storage and gradient computation. Before activation pruning our approach samples a probability of selecting a portion of parameters to freeze utilizing a bandit method for updates to prioritize impactful gradients on convergence. During the feed-forward pass each model layer adjusts adaptively based on parameter activation status obviating the need for sparsification and storage of redundant activations for subsequent backpropagation. Benchmarking on fine-tuning foundation models our approach maintains baseline accuracy across diverse tasks yielding over 20% speedup and around 10% memory reduction. Moreover integrating with an advanced CUDA kernel achieves up to 60% speedup without extra memory costs or accuracy loss significantly enhancing the efficiency of fine-tuning foundation models on memory-constrained devices.
****
Action-slot: Visual Action-centric Representations for Multi-label Atomic Activity Recognition in Traffic Scenes
Chi-Hsi Kung, Shu-Wei Lu, Yi-Hsuan Tsai, Yi-Ting Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18451-18461
In this paper we study multi-label atomic activity recognition. Despite the notable progress in action recognition it is still challenging to recognize atomic activities due to a deficiency in holistic understanding of both multiple road users' motions and their contextual information. In this paper we introduce Action-slot a slot attention-based approach that learns visual action-centric representations capturing both motion and contextual information. Our key idea is to design action slots that are capable of paying attention to regions where atomic activities occur without the need for explicit perception guidance. To further enhance slot attention we introduce a background slot that competes with action slots aiding the training process in avoiding unnecessary focus on background regions devoid of activities. Yet the imbalanced class distribution in the existing dataset hampers the assessment of rare activities. To address the limitation we collect a synthetic dataset called TACO which is four times larger than OATS and

features a balanced distribution of atomic activities. To validate the effective ness of our method we conduct comprehensive experiments and ablation studies aga inst various action recognition baselines. We also show that the performance of multi-label atomic activity recognition on real-world datasets can be improved b y pretraining representations on TACO.
****
Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Hu man Avatar Modeling
Zhe Li, Zerong Zheng, Lizhen Wang, Yebin Liu; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19711-19722
Modeling animatable human avatars from RGB videos is a long-standing and challen ging problem. Recent works usually adopt MLP-based neural radiance fields (NeRF) to represent 3D humans but it remains difficult for pure MLPs to regress pose-d ependent garment details. To this end we introduce Animatable Gaussians a new av atar representation that leverages powerful 2D CNNs and 3D Gaussian splatting to create high-fidelity avatars. To associate 3D Gaussians with the animatable ava tar we learn a parametric template from the input videos and then parameterize t he template on two front & back canonical Gaussian maps where each pixel represe nts a 3D Gaussian. The learned template is adaptive to the wearing garments for modeling looser clothes like dresses. Such template-guided 2D parameterization e nables us to employ a powerful StyleGAN-based CNN to learn the pose-dependent Ga ussian maps for modeling detailed dynamic appearances. Furthermore we introduce a pose projection strategy for better generalization given novel poses. Overall our method can create lifelike avatars with dynamic realistic and generalized ap pearances. Experiments show that our method outperforms other state-of-the-art a pproaches. Code: https://github.com/lizhe00/AnimatableGaussians.
****
Latency Correction for Event-guided Deblurring and Frame Interpolation
Yixin Yang, Jinxiu Liang, Bohan Yu, Yan Chen, Jimmy S. Ren, Boxin Shi; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24977-24986
Event cameras with their high temporal resolution dynamic range and low power co nsumption are particularly good at time-sensitive applications like deblurring a nd frame interpolation. However their performance is hindered by latency variabi lity especially under low-light conditions and with fast-moving objects. This pa per addresses the challenge of latency in event cameras -- the temporal discrepa ncy between the actual occurrence of changes in the corresponding timestamp assi gned by the sensor. Focusing on event-guided deblurring and frame interpolation tasks we propose a latency correction method based on a parameterized latency mo del. To enable data-driven learning we develop an event-based temporal fidelity to describe the sharpness of latent images reconstructed from events and the cor responding blurry images and reformulate the event-based double integral model d ifferentiable to latency. The proposed method is validated using synthetic and r eal-world datasets demonstrating the benefits of latency correction for deblurri ng and interpolation across different lighting conditions.
****
Retraining-Free Model Quantization via One-Shot Weight-Coupling Learning
Chen Tang, Yuan Meng, Jiacheng Jiang, Shuzhao Xie, Rongwei Lu, Xinzhu Ma, Zhi Wa ng, Wenwu Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2024, pp. 15855-15865
Quantization is of significance for compressing the over-parameterized deep neur al models and deploying them on resource-limited devices. Fixed-precision quanti zation suffers from performance drop due to the limited numerical representation ability. Conversely mixed-precision quantization (MPQ) is advocated to compress the model effectively by allocating heterogeneous bit-width for layers. MPQ is typically organized into a searching-retraining two-stage process. Previous work s only focus on determining the optimal bit-width configuration in the first sta ge efficiently while ignoring the considerable time costs in the second stage. H owever retraining always consumes hundreds of GPU-hours on the cutting-edge GPUs thus hindering deployment efficiency significantly. In this paper we devise a o

ne-shot training-searching paradigm for mixed-precision model compression. Speci
fically in the first stage all potential bit-width configurations are coupled an
d thus optimized simultaneously within a set of shared weights. However our obse
rvations reveal a previously unseen and severe bit-width interference phenomenon
 among highly coupled weights during optimization leading to considerable perfor
mance degradation under a high compression ratio. To tackle this problem we firs
t design a bit-width scheduler to dynamically freeze the most turbulent bit-widt
h of layers during training to ensure the rest bit-widths converged properly. Th
en taking inspiration from information theory we present an information distorti
on mitigation technique to align the behaviour of the bad-performing bit-widths
to the well-performing ones. In the second stage an inference-only greedy search
 scheme is devised to evaluate the goodness of configurations without introducin
g any additional training costs. Extensive experiments on three representative m
odels and three datasets demonstrate the effectiveness of the proposed method.
****
EVCap: Retrieval-Augmented Image Captioning with External Visual-Name Memory for
 Open-World Comprehension
Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, Hideki Nakayama; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
13733-13742
Large language models (LLMs)-based image captioning has the capability of descri
bing objects not explicitly observed in training data; yet novel objects occur f
requently necessitating the requirement of sustaining up-to-date object knowledg
e for open-world comprehension. Instead of relying on large amounts of data and/
or scaling up network parameters we introduce a highly effective retrieval-augme
nted image captioning method that prompts LLMs with object names retrieved from
External Visual--name memory (EVCap). We build ever-changing object knowledge me
mory using objects' visuals and names enabling us to (i) update the memory at a
minimal cost and (ii) effortlessly augment LLMs with retrieved object names by u
tilizing a lightweight and fast-to-train model. Our model which was trained only
 on the COCO dataset can adapt to out-of-domain without requiring additional fin
e-tuning or re-training. Our experiments conducted on benchmarks and synthetic c
ommonsense-violating data show that EVCap with only 3.97M trainable parameters e
xhibits superior performance compared to other methods based on frozen pre-train
ed LLMs. Its performance is also competitive to specialist SOTAs that require ex
tensive training.
****
SIFU: Side-view Conditioned Implicit Function for Real-world Usable Clothed Huma
n Reconstruction
Zechuan Zhang, Zongxin Yang, Yi Yang; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9936-9947
Creating high-quality 3D models of clothed humans from single images for real-wo
rld applications is crucial. Despite recent advancements accurately reconstructi
ng humans in complex poses or with loose clothing from in-the-wild images along
with predicting textures for unseen areas remains a significant challenge. A key
 limitation of previous methods is their insufficient prior guidance in transiti
oning from 2D to 3D and in texture prediction. In response we introduce SIFU (Si
de-view Conditioned Implicit Function for Real-world Usable Clothed Human Recons
truction) a novel approach combining a Side-view Decoupling Transformer with a 3
D Consistent Texture Refinement pipeline. SIFU employs a cross-attention mechani
sm within the transformer using SMPL-X normals as queries to effectively decoupl
e side-view features in the process of mapping 2D features to 3D. This method no
t only improves the precision of the 3D models but also their robustness especia
lly when SMPL-X estimates are not perfect. Our texture refinement process levera
ges text-to-image diffusion-based prior to generate realistic and consistent tex
tures for invisible views. Through extensive experiments SIFU surpasses SOTA met
hods in both geometry and texture reconstruction showcasing enhanced robustness
in complex scenarios and achieving an unprecedented Chamfer and P2S measurement.
 Our approach extends to practical applications such as 3D printing and scene bu
ilding demonstrating its broad utility in real-world scenarios.

****
WinSyn: : A High Resolution Testbed for Synthetic Data
Tom Kelly, John Femiani, Peter Wonka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22456-22465
We present WinSyn a unique dataset and testbed for creating high-quality synthetic data with procedural modeling techniques. The dataset contains high-resolution photographs of windows selected from locations around the world with 89318 individual window crops showcasing diverse geometric and material characteristics. We evaluate a procedural model by training semantic segmentation networks on both synthetic and real images and then comparing their performances on a shared test set of real images. Specifically we measure the difference in mean Intersection over Union (mIoU) and determine the effective number of real images to match synthetic data's training performance. We design a baseline procedural model as a benchmark and provide 21290 synthetically generated images. By tuning the procedural model key factors are identified which significantly influence the model's fidelity in replicating real-world scenarios. Importantly we highlight the challenge of procedural modeling using current techniques especially in their ability to replicate the spatial semantics of real-world scenarios. This insight is critical because of the potential of procedural models to bridge hidden scene aspects such as depth reflectivity material properties and lighting conditions.
****
Autoregressive Queries for Adaptive Tracking with Spatio-Temporal Transformers
Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19300-19309
The rich spatio-temporal information is crucial to capture the complicated target appearance variations in visual tracking. However most top-performing tracking algorithms rely on many hand-crafted components for spatio-temporal information aggregation. Consequently the spatio-temporal information is far away from being fully explored. To alleviate this issue we propose an adaptive tracker with spatio-temporal transformers (named AQATrack) which adopts simple autoregressive queries to effectively learn spatio-temporal information without many hand-designed components. Firstly we introduce a set of learnable and autoregressive queries to capture the instantaneous target appearance changes in a sliding window fashion. Then we design a novel attention mechanism for the interaction of existing queries to generate a new query in current frame. Finally based on the initial target template and learnt autoregressive queries a spatio-temporal information fusion module (STM) is designed for spatiotemporal formation aggregation to locate a target object. Benefiting from the STM we can effectively combine the static appearance and instantaneous changes to guide robust tracking. Extensive experiments show that our method significantly improves the tracker's performance on six popular tracking benchmarks: LaSOT LaSOText TrackingNet GOT-10k TNL2K and UAV123. Code and models will be https://github.com/orgs/GXNU-ZhongLab.
****
Misalignment-Robust Frequency Distribution Loss for Image Transformation
Zhangkai Ni, Juncheng Wu, Zian Wang, Wenhan Yang, Hanli Wang, Lin Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2910-2919
This paper aims to address a common challenge in deep learning-based image transformation methods such as image enhancement and super-resolution which heavily rely on precisely aligned paired datasets with pixel-level alignments. However creating precisely aligned paired images presents significant challenges and hinders the advancement of methods trained on such data. To overcome this challenge this paper introduces a novel and simple Frequency Distribution Loss (FDL) for computing distribution distance within the frequency domain. Specifically we transform image features into the frequency domain using Discrete Fourier Transformation (DFT). Subsequently frequency components (amplitude and phase) are processed separately to form the FDL loss function. Our method is empirically proven effective as a training constraint due to the thoughtful utilization of global information in the frequency domain. Extensive experimental evaluations focusing on i

mage enhancement and super-resolution tasks demonstrate that FDL outperforms existing misalignment-robust loss functions. Furthermore we explore the potential of our FDL for image style transfer that relies solely on completely misaligned data. Our code is available at: https://github.com/eezkni/FDL
****
Language-aware Visual Semantic Distillation for Video Question Answering
Bo Zou, Chao Yang, Yu Qiao, Chengbin Quan, Youjian Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27113-27123
Significant advancements in video question answering (VideoQA) have been made thanks to thriving large image-language pretraining frameworks. Although these image-language models can efficiently represent both video and language branches they typically employ a goal-free vision perception process and do not interact vision with language well during the answer generation thus omitting crucial visual cues. In this paper we are inspired by the human recognition and learning pattern and propose VideoDistill a framework with language-aware (i.e. goal-driven) behavior in both vision perception and answer generation process. VideoDistill generates answers only from question-related visual embeddings and follows a thinking-observing-answering approach that closely resembles human behavior distinguishing it from previous research. Specifically we develop a language-aware gating mechanism to replace the standard cross-attention avoiding language's direct fusion into visual representations. We incorporate this mechanism into two key components of the entire framework. The first component is a differentiable sparse sampling module which selects frames containing the necessary dynamics and semantics relevant to the questions. The second component is a vision refinement module that merges existing spatial-temporal attention layers to ensure the extraction of multi-grained visual semantics associated with the questions. We conduct experimental evaluations on various challenging video question-answering benchmarks and VideoDistill achieves state-of-the-art performance in both general and long-form VideoQA datasets. In Addition we verify that VideoDistill can effectively alleviate the utilization of language shortcut solutions in the EgoTaskQA dataset.
****
Lane2Seq: Towards Unified Lane Detection via Sequence Generation
Kunyang Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16944-16953
In this paper we present a novel sequence generation-based framework for lane detection called Lane2Seq. It unifies various lane detection formats by casting lane detection as a sequence generation task. This is different from previous lane detection methods which depend on well-designed task-specific head networks and corresponding loss functions. Lane2Seq only adopts a plain transformer-based encoder-decoder architecture with a simple cross-entropy loss. Additionally we propose a new multi-format model tuning based on reinforcement learning to incorporate the task-specific knowledge into Lane2Seq. Experimental results demonstrate that such a simple sequence generation paradigm not only unifies lane detection but also achieves competitive performance on benchmarks. For example Lane2Seq gets 97.95% and 97.42% F1 score on Tusimple and LLAMAS datasets establishing a new state-of-the-art result for two benchmarks.
****
Disentangled Prompt Representation for Domain Generalization
De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, Xinbo Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23595-23604
Domain Generalization (DG) aims to develop a versatile model capable of performing well on unseen target domains. Recent advancements in pre-trained Visual Foundation Models (VFMs) such as CLIP show significant potential in enhancing the generalization abilities of deep models. Although there is a growing focus on VFM-based domain prompt tuning for DG effectively learning prompts that disentangle invariant features across all domains remains a major challenge. In this paper we propose addressing this challenge by leveraging the controllable and flexible

language prompt of the VFM. Observing that the text modality of VFMs is inherent
ly easier to disentangle we introduce a novel text feature guided visual prompt
tuning framework. This framework first automatically disentangles the text promp
t using a large language model (LLM) and then learns domain-invariant visual rep
resentation guided by the disentangled text feature. Moreover we also devise dom
ain-specific prototype learning to fully exploit domain-specific information to
combine with the invariant feature prediction. Extensive experiments on mainstre
am DG datasets namely PACS VLCS OfficeHome DomainNet and TerraInc demonstrate th
at the proposed method achieves superior performances to state-of-the-art DG met
hods.
****
Abductive Ego-View Accident Video Understanding for Safe Driving Perception
Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru X
ue, Tat-Seng Chua; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2024, pp. 22030-22040
We present MM-AU a novel dataset for Multi-Modal Accident video Understanding. M
M-AU contains 11727 in-the-wild ego-view accident videos each with temporally al
igned text descriptions. We annotate over 2.23 million object boxes and 58650 pa
irs of video-based accident reasons covering 58 accident categories. MM-AU suppo
rts various accident understanding tasks particularly multimodal video diffusion
 to understand accident cause-effect chains for safe driving. With MM-AU we pres
ent an Abductive accident Video understanding framework for Safe Driving percept
ion (AdVersa-SD). AdVersa-SD performs video diffusion via an Object-Centric Vide
o Diffusion (OAVD) method which is driven by an abductive CLIP model. This model
 involves a contrastive interaction loss to learn the pair co-occurrence of norm
al near-accident accident frames with the corresponding text descriptions such a
s accident reasons prevention advice and accident categories. OAVD enforces the
object region learning while fixing the content of the original frame background
 in video generation to find the dominant objects for certain accidents. Extensi
ve experiments verify the abductive ability of AdVersa-SD and the superiority of
 OAVD against the state-of-the-art diffusion models. Additionally we provide car
eful benchmark evaluations for object detection and accident reason answering si
nce AdVersa-SD relies on precise object and accident reason information.
****
Cross-spectral Gated-RGB Stereo Depth Estimation
Samuel Brucker, Stefanie Walz, Mario Bijelic, Felix Heide; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2
1654-21665
Gated cameras flood-illuminate a scene and capture the time-gated impulse respon
se of a scene. By employing nanosecond-scale gates existing sensors are capable
of capturing mega-pixel gated images delivering dense depth improving on today's
 LiDAR sensors in spatial resolution and depth precision. Although gated depth e
stimation methods deliver a million of depth estimates per frame their resolutio
n is still an order below existing RGB imaging methods. In this work we combine
high-resolution stereo HDR RCCB cameras with gated imaging allowing us to exploi
t depth cues from active gating multi-view RGB and multi-view NIR sensing -- mul
ti-view and gated cues across the entire spectrum. The resulting capture system
consists only of low-cost CMOS sensors and flood-illumination. We propose a nove
l stereo-depth estimation method that is capable of exploiting these multi-modal
 multi-view depth cues including the active illumination that is measured by the
 RCCB camera when removing the IR-cut filter. The proposed method achieves accur
ate depth at long ranges outperforming the next best existing method by 39% for
ranges of 100 to 220 m in MAE on accumulated LiDAR ground-truth. Our code models
 and datasets are available here (https://light.princeton.edu/gatedrccbstereo/).
****
KVQ: Kwai Video Quality Assessment for Short-form Videos
Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Z
hou, Zhibo Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2024, pp. 25963-25973
Short-form UGC video platforms like Kwai and TikTok have been an emerging and ir

replaceable mainstream media form thriving on user-friendly engagement and kalei doscope creation etc. However the advancing content generation modes e.g. specia l effects and sophisticated processing workflows e.g. de-artifacts have introduc ed significant challenges to recent UGC video quality assessment: (i) the ambigu ous contents hinder the identification of quality-determined regions. (ii) the d iverse and complicated hybrid distortions are hard to distinguish. To tackle the above challenges and assist in the development of short-form videos we establis h the first large-scale Kwai short Video database for Quality assessment termed KVQ which comprises 600 user-uploaded short videos and 3600 processed videos thr ough the diverse practical processing workflows including pre-processing transco ding and enhancement. Among them the absolute quality score of each video and pa rtial ranking score among indistinguish samples are provided by a team of profes sional researchers specializing in image processing. Based on this database we p ropose the first short-form video quality evaluator i.e. KSVQE which enables the quality evaluator to identify the quality-determined semantics with the content understanding of large vision language models (i.e. CLIP) and distinguish the d istortions with the distortion under- standing module. Experimental results have shown the effectiveness of KSVQE on our KVQ database and popular VQA databases. The project can be found at https: //lixinustc.github.io/projects/KVQ/.
****
Degrees of Freedom Matter: Inferring Dynamics from Point Trajectories
Yan Zhang, Sergey Prokudin, Marko Mihajlovic, Qianli Ma, Siyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 24, pp. 2018-2028
Understanding the dynamics of generic 3D scenes is fundamentally challenging in computer vision essential in enhancing applications related to scene reconstruct ion motion tracking and avatar creation. In this work we address the task as the problem of inferring dense long-range motion of 3D points. By observing a set o f point trajectories we aim to learn an implicit motion field parameterized by a neural network to predict the movement of novel points within the same domain w ithout relying on any data-driven or scene-specific priors. To achieve this our approach builds upon the recently introduced dynamic point field model that lear ns smooth deformation fields between the canonical frame and individual observat ion frames. However temporal consistency between consecutive frames is neglected and the number of required parameters increases linearly with the sequence leng th due to per-frame modeling. To address these shortcomings we exploit the intri nsic regularization provided by SIREN and modify the input layer to produce a sp atiotemporally smooth motion field. Additionally we analyze the motion field Jac obian matrix and discover that the motion degrees of freedom (DOFs) in an infini tesimal area around a point and the network hidden variables have different beha viors to affect the model's representational power. This enables us to improve t he model representation capability while retaining the model compactness. Furthe rmore to reduce the risk of overfitting we introduce a regularization term based on the assumption of piece-wise motion smoothness. Our experiments assess the m odel's performance in predicting unseen point trajectories and its application i n temporal mesh alignment with guidance. The results demonstrate its superiority and effectiveness. The code and data for the project are publicly available at https://yz-cnsdqz.github.io/eigenmotion/DOMA.
****
LEMON: Learning 3D Human-Object Interaction Relation from 2D Images
Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp . 16284-16295
Learning 3D human-object interaction relation is pivotal to embodied AI and inte raction modeling. Most existing methods approach the goal by learning to predict isolated interaction elements e.g. human contact object affordance and human-ob ject spatial relation primarily from the perspective of either the human or the object. Which underexploit certain correlations between the interaction counterp arts (human and object) and struggle to address the uncertainty in interactions. Actually objects' functionalities potentially affect humans' interaction intent

ions which reveals what the interaction is. Meanwhile the interacting humans and objects exhibit matching geometric structures which presents how to interact. In light of this we propose harnessing these inherent correlations between interaction counterparts to mitigate the uncertainty and jointly anticipate the above interaction elements in 3D space. To achieve this we present LEMON (LEarning 3D huMan-Object iNteraction relation) a unified model that mines interaction intentions of the counterparts and employs curvatures to guide the extraction of geometric correlations combining them to anticipate the interaction elements. Besides the 3D Interaction Relation dataset (3DIR) is collected to serve as the test bed for training and evaluation. Extensive experiments demonstrate the superiority of LEMON over methods estimating each element in isolation. The code and dataset are available at https://yyvhang.github.io/LEMON/
****
Low-Latency Neural Stereo Streaming
Qiqi Hou, Farzad Farhadzadeh, Amir Said, Guillaume Sautiere, Hoang Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7974-7984
The rise of new video modalities like virtual reality or autonomous driving has increased the demand for efficient multi-view video compression methods both in terms of rate-distortion (R-D) performance and in terms of delay and runtime. While most recent stereo video compression approaches have shown promising performance they compress left and right views sequentially leading to poor parallelization and runtime performance. This work presents Low-Latency neural codec for Stereo video Streaming (LLSS) a novel parallel stereo video coding method designed for fast and efficient low-latency stereo video streaming. Instead of using a sequential cross-view motion compensation like existing methods LLSS introduces a bidirectional feature shifting module to directly exploit mutual information among views and encode them effectively with a joint cross-view prior model for entropy coding. Thanks to this design LLSS processes left and right views in parallel minimizing latency; all while substantially improving R-D performance compared to both existing neural and conventional codecs.
****
Understanding Video Transformers via Universal Concept Discovery
Matthew Kowal, Achal Dave, Rares Ambrus, Adrien Gaidon, Konstantinos G. Derpanis, Pavel Tokmakov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10946-10956
This paper studies the problem of concept-based interpretability of transformer representations for videos. Concretely we seek to explain the decision-making process of video transformers based on high-level spatiotemporal concepts that are automatically discovered. Prior research on concept-based interpretability has concentrated solely on image-level tasks. Comparatively video models deal with the added temporal dimension increasing complexity and posing challenges in identifying dynamic concepts over time. In this work we systematically address these challenges by introducing the first Video Transformer Concept Discovery (VTCD) algorithm. To this end we propose an efficient approach for unsupervised identification of units of video transformer representations - concepts and ranking their importance to the output of a model. The resulting concepts are highly interpretable revealing spatio-temporal reasoning mechanisms and object-centric representations in unstructured video models. Performing this analysis jointly over a diverse set of supervised and self-supervised representations we discover that some of these mechanism are universal in video transformers. Finally we show that VTCD can be used for fine-grained action recognition and video object segmentation.
****
Exploring the Transferability of Visual Prompting for Multimodal Large Language Models
Yichi Zhang, Yinpeng Dong, Siyuan Zhang, Tianzan Min, Hang Su, Jun Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26562-26572
Although Multimodal Large Language Models (MLLMs) have demonstrated promising ve

rsatile capabilities their performance is still inferior to specialized models o
n downstream tasks which makes adaptation necessary to enhance their utility. Ho
wever fine-tuning methods require independent training for every model leading t
o huge computation and memory overheads. In this paper we propose a novel settin
g where we aim to improve the performance of diverse MLLMs with a group of share
d parameters optimized for a downstream task. To achieve this we propose Transfe
rable Visual Prompting (TVP) a simple and effective approach to generate visual
prompts that can transfer to different models and improve their performance on d
ownstream tasks after trained on only one model. We introduce two strategies to
address the issue of cross-model feature corruption of existing visual prompting
 methods and enhance the transferability of the learned prompts including 1) Fea
ture Consistency Alignment: which imposes constraints to the prompted feature ch
anges to maintain task-agnostic knowledge; 2) Task Semantics Enrichment: which e
ncourages the prompted images to contain richer task-specific semantics with lan
guage guidance. We validate the effectiveness of TVP through extensive experimen
ts with 6 modern MLLMs on a wide variety of tasks ranging from object recognitio
n and counting to multimodal reasoning and hallucination correction.
****
PointOBB: Learning Oriented Object Detection via Single Point Supervision
Junwei Luo, Xue Yang, Yi Yu, Qingyun Li, Junchi Yan, Yansheng Li; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 16730-16740
Single point-supervised object detection is gaining attention due to its cost-ef
fectiveness. However existing approaches focus on generating horizontal bounding
 boxes (HBBs) while ignoring oriented bounding boxes (OBBs) commonly used for ob
jects in aerial images. This paper proposes PointOBB the first single Point-base
d OBB generation method for oriented object detection. PointOBB operates through
 the collaborative utilization of three distinctive views: an original view a re
sized view and a rotated/flipped (rot/flp) view. Upon the original view we lever
age the resized and rot/flp views to build a scale augmentation module and an an
gle acquisition module respectively. In the former module a Scale-Sensitive Cons
istency (SSC) loss is designed to enhance the deep network's ability to perceive
 the object scale. For accurate object angle predictions the latter module incor
porates self-supervised learning to predict angles which is associated with a sc
ale-guided Dense-to-Sparse (DS) matching strategy for aggregating dense angles c
orresponding to sparse objects. The resized and rot/flp views are switched using
 a progressive multi-view switching strategy during training to achieve coupled
optimization of scale and angle. Experimental results on the DIOR-R and DOTA-v1.
0 datasets demonstrate that PointOBB achieves promising performance and signific
antly outperforms potential point-supervised baselines.
****
Intrinsic Image Diffusion for Indoor Single-view Material Estimation
Peter Kocsis, Vincent Sitzmann, Matthias Nießner; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5198-5208
We present Intrinsic Image Diffusion a generative model for appearance decomposi
tion of indoor scenes. Given a single input view we sample multiple possible mat
erial explanations represented as albedo roughness and metallic maps. Appearance
 decomposition poses a considerable challenge in computer vision due to the inhe
rent ambiguity between lighting and material properties and the lack of real dat
asets. To address this issue we advocate for a probabilistic formulation where i
nstead of attempting to directly predict the true material properties we employ
a conditional generative model to sample from the solution space. Furthermore we
 show that utilizing the strong learned prior of recent diffusion models trained
 on large-scale real-world images can be adapted to material estimation and high
ly improves the generalization to real images. Our method produces significantly
 sharper more consistent and more detailed materials outperforming state-of-the-
art methods by 1.5dB on PSNR and by 45% better FID score on albedo prediction. W
e demonstrate the effectiveness of our approach through experiments on both synt
hetic and real-world datasets.
****

SHAP-EDITOR: Instruction-Guided Latent 3D Editing in Seconds
Minghao Chen, Junyu Xie, Iro Laina, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26456-26466
We propose a novel feed-forward 3D editing framework called Shap-Editor. Prior research on editing 3D objects primarily concentrated on editing individual objects by leveraging off-the-shelf 2D image editing networks utilizing a process called 3D distillation which transfers knowledge from the 2D network to the 3D asset. Distillation necessitates at least tens of minutes per asset to attain satisfactory editing results thus it is not very practical. In contrast we ask whether 3D editing can be carried out directly by a feed-forward network eschewing test-time optimization. In particular we hypothesise that this process can be greatly simplified by first encoding 3D objects into a suitable latent space. We validate this hypothesis by building upon the latent space of Shap-E. We demonstrate that direct 3D editing in this space is possible and efficient by learning a feed-forward editor network that only requires approximately one second per edit. Our experiments show that Shap-Editor generalises well to both in-distribution and out-of-distribution 3D assets with different prompts and achieves superior performance compared to methods that carry out test-time optimisation for each edited instance.
****
HyperSDFusion: Bridging Hierarchical Structures in Language and Geometry for Enhanced 3D Text2Shape Generation
Zhiying Leng, Tolga Birdal, Xiaohui Liang, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19691-19700
3D shape generation from text is a fundamental task in 3D representation learning. The text-shape pairs exhibit a hierarchical structure where a general text like "chair" covers all 3D shapes of the chair while more detailed prompts refer to more specific shapes. Furthermore both text and 3D shapes are inherently hierarchical structures. However existing Text2Shape methods such as SDFusion do not exploit that. In this work we propose HyperSDFusion a dual-branch diffusion model that generates 3D shapes from a given text. Since hyperbolic space is suitable for handling hierarchical data we propose to learn the hierarchical representations of text and 3D shapes in hyperbolic space. First we introduce a hyperbolic text-image encoder to learn the sequential and multi-modal hierarchical features of text in hyperbolic space. In addition we design a hyperbolic text-graph convolution module to learn the hierarchical features of text in hyperbolic space. In order to fully utilize these text features we introduce a dual-branch structure to embed text features in 3D feature space. At last to endow the generated 3D shapes with a hierarchical structure we devise a hyperbolic hierarchical loss. Our method is the first to explore the hyperbolic hierarchical representation for text-to-shape generation. Experimental results on the existing text-to-shape paired dataset Text2Shape achieved state-of-the-art results. We release our implementation under HyperSDFusion.github.io.
****
OmniParser: A Unified Framework for Text Spotting Key Information Extraction and Table Recognition
Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, Zhibo Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15641-15653
Recently visually-situated text parsing (VsTP) has experienced notable advancements driven by the increasing demand for automated document understanding and the emergence of Generative Large Language Models (LLMs) capable of processing document-based questions. Various methods have been proposed to address the challenging problem of VsTP. However due to the diversified targets and heterogeneous schemas previous works usually design task-specific architectures and objectives for individual tasks which inadvertently leads to modal isolation and complex workflow. In this paper we propose a unified paradigm for parsing visually-situated text across diverse scenarios. Specifically we devise a universal model called

OmniParser which can simultaneously handle three typical visually-situated text parsing tasks: text spotting key information extraction and table recognition. In OmniParser all tasks share the unified encoder-decoder architecture the unified objective: point-conditioned text generation and the unified input & output representation: prompt & structured sequences. Extensive experiments demonstrate that the proposed OmniParser achieves state-of-the-art (SOTA) or highly competitive performances on 7 datasets for the three visually-situated text parsing tasks despite its unified concise design. The code is available at https://github.com/AlibabaResearch/AdvancedLiterateMachinery.
****
Are Conventional SNNs Really Efficient? A Perspective from Network Quantization
Guobin Shen, Dongcheng Zhao, Tenglong Li, Jindong Li, Yi Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27538-27547
Spiking Neural Networks (SNNs) have been widely praised for their high energy efficiency and immense potential. However comprehensive research that critically contrasts and correlates SNNs with quantized Artificial Neural Networks (ANNs) remains scant often leading to skewed comparisons lacking fairness towards ANNs. This paper introduces a unified perspective illustrating that the time steps in SNNs and quantized bit-widths of activation values present analogous representations. Building on this we present a more pragmatic and rational approach to estimating the energy consumption of SNNs. Diverging from the conventional Synaptic Operations (SynOps) we champion the "Bit Budget" concept. This notion permits an intricate discourse on strategically allocating computational and storage resources between weights activation values and temporal steps under stringent hardware constraints. Guided by the Bit Budget paradigm we discern that pivoting efforts towards spike patterns and weight quantization rather than temporal attributes elicits profound implications for model performance. Utilizing the Bit Budget for holistic design consideration of SNNs elevates model performance across diverse data types encompassing static imagery and neuromorphic datasets. Our revelations bridge the theoretical chasm between SNNs and quantized ANNs and illuminate a pragmatic trajectory for future endeavors in energy-efficient neural computations.
****
Training Like a Medical Resident: Context-Prior Learning Toward Universal Medical Image Segmentation
Yunhe Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11194-11204
A major focus of clinical imaging workflow is disease diagnosis and management leading to medical imaging datasets strongly tied to specific clinical objectives. This scenario has led to the prevailing practice of developing task-specific segmentation models without gaining insights from widespread imaging cohorts. Inspired by the training program of medical radiology residents we propose a shift towards universal medical image segmentation a paradigm aiming to build medical image understanding foundation models by leveraging the diversity and commonality across clinical targets body regions and imaging modalities. Towards this goal we develop Hermes a novel context-prior learning approach to address the challenges of data heterogeneity and annotation differences in medical image segmentation. In a large collection of eleven diverse datasets (2438 3D images) across five modalities (CT PET T1 T2 and cine MRI) and multiple body regions we demonstrate the merit of the universal paradigm over the traditional paradigm on addressing multiple tasks within a single model. By exploiting the synergy across tasks Hermes achieves state-of-the-art performance on all testing datasets and shows superior model scalability. Results on two additional datasets reveals Hermes' strong performance for transfer learning incremental learning and generalization to downstream tasks. Hermes's learned priors demonstrate an appealing trait to reflect the intricate relations among tasks and modalities which aligns with the established anatomical and imaging principles in radiology. The code is available.
****

Material Palette: Extraction of Materials from a Single Image
Ivan Lopes, Fabio Pizzati, Raoul de Charette; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4379-4388
Physically-Based Rendering (PBR) is key to modeling the interaction between ligh t and materials and finds extensive applications across computer graphics domain s. However acquiring PBR materials is costly and requires special apparatus. In this paper we propose a method to extract PBR materials from a single real-world image. We do so in two steps: first we map regions of the image to material con cept tokens using a diffusion model allowing the sampling of texture images rese mbling each material in the scene. Second we leverage a separate network to deco mpose the generated textures into spatially varying BRDFs (SVBRDFs) offering us readily usable materials for rendering applications. Our approach relies on exis ting synthetic material libraries with SVBRDF ground truth. It exploits a diffus ion-generated RGB texture dataset to allow generalization to new samples using u nsupervised domain adaptation (UDA). Our contributions are thoroughly evaluated on synthetic and real-world datasets. We further demonstrate the applicability o f our method for editing 3D scenes with materials estimated from real photograph s. Along with video we share code and models as open-source on the project page: https://github.com/astra-vision/MaterialPalette
****
Initialization Matters for Adversarial Transfer Learning
Andong Hua, Jindong Gu, Zhiyu Xue, Nicholas Carlini, Eric Wong, Yao Qin; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24831-24840
With the prevalence of the Pretraining-Finetuning paradigm in transfer learning the robustness of downstream tasks has become a critical concern. In this work w e delve into adversarial robustness in transfer learning and reveal the critical role of initialization including both the pretrained model and the linear head. First we discover the necessity of an adversarially robust pretrained model. Sp ecifically we reveal that with a standard pretrained model Parameter-Efficient F inetuning (PEFT) methods either fail to be adversarially robust or continue to e xhibit significantly degraded adversarial robustness on downstream tasks even wi th adversarial training during finetuning. Leveraging a robust pretrained model surprisingly we observe that a simple linear probing can outperform full finetun ing and other PEFT methods with random initialization on certain datasets. We fu rther identify that linear probing excels in preserving robustness from the robu st pretraining. Based on this we propose Robust Linear Initialization (RoLI) for adversarial finetuning which initializes the linear head with the weights obtai ned by adversarial linear probing to maximally inherit the robustness from pretr aining. Across five different image classification datasets we demonstrate the e ffectiveness of RoLI and achieve new state-of-the-art results. Our code is avail able at https://github.com/DongXzz/RoLI.
****
RealCustom: Narrowing Real Text Word for Real-Time Open-Domain Text-to-Image Cus tomization
Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, Yongdong Zhang; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 4, pp. 7476-7485
Text-to-image customization which aims to synthesize text-driven images for the given subjects has recently revolutionized content creation. Existing works foll ow the pseudo-word paradigm i.e. represent the given subjects as pseudo-words an d then compose them with the given text. However the inherent entangled influenc e scope of pseudo-words with the given text results in a dual-optimum paradox i. e. the similarity of the given subjects and the controllability of the given tex t could not be optimal simultaneously. We present RealCustom that for the first time disentangles similarity from controllability by precisely limiting subject influence to relevant parts only achieved by gradually narrowing real text word from its general connotation to the specific subject and using its cross-attenti on to distinguish relevance. Specifically RealCustom introduces a novel "train-i nference" decoupled framework: (1) during training RealCustom learns general ali

gnment between visual conditions to original textual conditions by a novel adapt
ive scoring module to adaptively modulate influence quantity; (2) during inferen
ce a novel adaptive mask guidance strategy is proposed to iteratively update the
 influence scope and influence quantity of the given subjects to gradually narro
w the generation of the real text word. Comprehensive experiments demonstrate th
e superior real-time customization ability of RealCustom in the open domain achi
eving both unprecedented similarity of the given subjects and controllability of
 the given text for the first time. The project page is https://corleone-huang.g
ithub.io/realcustom/.
****
MicroDiffusion: Implicit Representation-Guided Diffusion for 3D Reconstruction f
rom Limited 2D Microscopy Projections
Mude Hui, Zihao Wei, Hongru Zhu, Fei Xia, Yuyin Zhou; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11460-
11469
Volumetric optical microscopy using non-diffracting beams enables rapid imaging
of 3D volumes by projecting them axially to 2D images but lacks crucial depth in
formation. Addressing this we introduce MicroDiffusion a pioneering tool facilit
ating high-quality depth-resolved 3D volume reconstruction from limited 2D proje
ctions. While existing Implicit Neural Representation (INR) models often yield i
ncomplete outputs and Denoising Diffusion Probabilistic Models (DDPM) excel at c
apturing details our method integrates INR's structural coherence with DDPM's fi
ne-detail enhancement capabilities. We pretrain an INR model to transform 2D axi
ally-projected images into a preliminary 3D volume. This pretrained INR acts as
a global prior guiding DDPM's generative process through a linear interpolation
between INR outputs and noise inputs. This strategy enriches the diffusion proce
ss with structured 3D information enhancing detail and reducing noise in localiz
ed 2D images.By conditioning the diffusion model on the closest 2D projection Mi
croDiffusion substantially enhances fidelity in resulting 3D reconstructions sur
passing INR and standard DDPM outputs with unparalleled image quality and struct
ural fidelity. Our code and dataset are available athttps://github.com/UCSC-VLAA
/MicroDiffusion.
****
Task-Conditioned Adaptation of Visual Features in Multi-Task Policy Learning
Pierre Marza, Laetitia Matignon, Olivier Simonin, Christian Wolf; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
, pp. 17847-17856
Successfully addressing a wide variety of tasks is a core ability of autonomous
agents requiring flexibly adapting the underlying decision-making strategies and
 as we argue in this work also adapting the perception modules. An analogical ar
gument would be the human visual system which uses top-down signals to focus att
ention determined by the current task. Similarly we adapt pre-trained large visi
on models conditioned on specific downstream tasks in the context of multi-task
policy learning. We introduce task-conditioned adapters that do not require fine
tuning any pre-trained weights combined with a single policy trained with behavi
or cloning and capable of addressing multiple tasks. We condition the visual ada
pters on task embeddings which can be selected at inference if the task is known
 or alternatively inferred from a set of example demonstrations. To this end we
propose a new optimization-based estimator. We evaluate the method on a wide var
iety of tasks from the CortexBench benchmark and show that compared to existing
work it can be addressed with a single policy. In particular we demonstrate that
 adapting visual features is a key design choice and that the method generalizes
 to unseen tasks given a few demonstrations.
****
L0-Sampler: An L0 Model Guided Volume Sampling for NeRF
Liangchen Li, Juyong Zhang; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2024, pp. 21390-21400
Since its proposal Neural Radiance Fields (NeRF) has achieved great success in r
elated tasks mainly adopting the hierarchical volume sampling (HVS) strategy for
 volume rendering. However the HVS of NeRF approximates distributions using piec

ewise constant functions which provides a relatively rough estimation. Based on the observation that a well-trained weight function w(t) and the L_0 distance between points and the surface have very high similarity we propose L_0-Sampler by incorporating the L_0 model into w(t) to guide the sampling process. Specifically we propose using piecewise exponential functions rather than piecewise constant functions for interpolation which can not only approximate quasi-L_0 weight distributions along rays quite well but can be easily implemented with a few lines of code change without additional computational burden. Stable performance improvements can be achieved by applying L_0-Sampler to NeRF and related tasks like 3D reconstruction. Code is available at https://ustc3dv.github.io/L0-Sampler/.
****
Hybrid Proposal Refiner: Revisiting DETR Series from the Faster R-CNN Perspective
Jinjing Zhao, Fangyun Wei, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17416-17426
With the transformative impact of the Transformer DETR pioneered the application of the encoder-decoder architecture to object detection. A collection of follow-up research e.g. Deformable DETR aims to enhance DETR while adhering to the encoder-decoder design. In this work we revisit the DETR series through the lens of Faster R-CNN. We find that the DETR resonates with the underlying principles of Faster R-CNN's RPN-refiner design but benefits from end-to-end detection owing to the incorporation of Hungarian matching. We systematically adapt the Faster R-CNN towards the Deformable DETR by integrating or repurposing each component of Deformable DETR and note that Deformable DETR's improved performance over Faster R-CNN is attributed to the adoption of advanced modules such as a superior proposal refiner (e.g. deformable attention rather than RoI Align). When viewing the DETR through the RPN-refiner paradigm we delve into various proposal refinement techniques such as deformable attention cross attention and dynamic convolution. These proposal refiners cooperate well with each other; thus we synergistically combine them to establish a Hybrid Proposal Refiner (HPR). Our HPR is versatile and can be incorporated into various DETR detectors. For instance by integrating HPR to a strong DETR detector we achieve an AP of 54.9 on the COCO benchmark utilizing a ResNet-50 backbone and a 36-epoch training schedule. Code and models are available at https://github.com/ZhaoJingjing713/HPR.
****
Practical Measurements of Translucent Materials with Inter-Pixel Translucency Prior
Zhenyu Chen, Jie Guo, Shuichang Lai, Ruoyu Fu, Mengxun Kong, Chen Wang, Hongyu Sun, Zhebin Zhang, Chen Li, Yanwen Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20932-20942
Material appearance is a key component of photorealism with a pronounced impact on human perception. Although there are many prior works targeting at measuring opaque materials using light-weight setups (e.g. consumer-level cameras) little attention is paid on acquiring the optical properties of translucent materials which are also quite common in nature. In this paper we present a practical method for acquiring scattering properties of translucent materials based solely on ordinary images captured with unknown lighting and camera parameters. The key to our method is an inter-pixel translucency prior which states that image pixels of a given homogeneous translucent material typically form curves (dubbed translucent curves) in the RGB space of which the shapes are determined by the parameters of the material. We leverage this prior in a specially-designed convolutional neural network comprising multiple encoders a translucency-aware feature fusion module and a cascaded decoder. We demonstrate through both visual comparisons and quantitative evaluations that high accuracy can be achieved on a wide range of real-world translucent materials.
****
TurboSL: Dense Accurate and Fast 3D by Neural Inverse Structured Light
Parsa Mirdehghan, Maxx Wu, Wenzheng Chen, David B. Lindell, Kiriakos N. Kutulakos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25067-25076

We show how to turn a noisy and fragile active triangulation technique--three-pa
ttern structured light with a grayscale camera--into a fast and powerful tool fo
r 3D capture: able to output sub-pixel accurate disparities at megapixel resolut
ion along with reflectance normals and a no-reference estimate of its own pixelw
ise 3D error. To achieve this we formulate structured-light decoding as a neural
 inverse rendering problem. We show that despite having just three or four input
 images--all from the same viewpoint--this problem can be tractably solved by Tu
rboSL an algorithm that combines (1) a precise image formation model (2) a signe
d distance field scene representation and (3) projection pattern sequences optim
ized for accuracy instead of precision. We use TurboSL to reconstruct a variety
of complex scenes from images captured at up to 60 fps with a camera and a commo
n projector. Our experiments highlight TurboSL's potential for dense and highly-
accurate 3D acquisition from data captured in fractions of a second.
****
Text2QR: Harmonizing Aesthetic Customization and Scanning Robustness for Text-Gu
ided QR Code Generation
Guangyang Wu, Xiaohong Liu, Jun Jia, Xuehao Cui, Guangtao Zhai; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 8456-8465
In the digital era QR codes serve as a linchpin connecting virtual and physical
realms. Their pervasive integration across various applications highlights the d
emand for aesthetically pleasing codes without compromised scannability. However
 prevailing methods grapple with the intrinsic challenge of balancing customizat
ion and scannability. Notably stable-diffusion models have ushered in an epoch o
f high-quality customizable content generation. This paper introduces Text2QR a
pioneering approach leveraging these advancements to address a fundamental chall
enge: concurrently achieving user-defined aesthetics and scanning robustness. To
 ensure stable generation of aesthetic QR codes we introduce the QR Aesthetic Bl
ueprint (QAB) module generating a blueprint image exerting control over the enti
re generation process. Subsequently the Scannability Enhancing Latent Refinement
 (SELR) process refines the output iteratively in the latent space enhancing sca
nning robustness. This approach harnesses the potent generation capabilities of
stable-diffusion models navigating the trade-off between image aesthetics and QR
 code scannability. Our experiments demonstrate the seamless fusion of visual ap
peal with the practical utility of aesthetic QR codes markedly outperforming pri
or methods. Codes are available at https://github.com/mulns/Text2QR
****
GS-IR: 3D Gaussian Splatting for Inverse Rendering
Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, Kui Jia; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21644
-21653
We propose GS-IR a novel inverse rendering approach based on 3D Gaussian Splatti
ng (GS) that leverages forward mapping volume rendering to achieve photorealisti
c novel view synthesis and relighting results. Unlike previous works that use im
plicit neural representations and volume rendering (e.g. NeRF) which suffer from
 low expressive power and high computational complexity we extend GS a top-perfo
rmance representation for novel view synthesis to estimate scene geometry surfac
e material and environment illumination from multi-view images captured under un
known lighting conditions. There are two main problems when introducing GS to in
verse rendering: 1) GS does not support producing plausible normal natively; 2)
forward mapping (e.g. rasterization and splatting) cannot trace the occlusion li
ke backward mapping (e.g. ray tracing). To address these challenges our GS-IR pr
oposes an efficient optimization scheme that incorporates a depth-derivation-bas
ed regularization for normal estimation and a baking-based occlusion to model in
direct lighting. The flexible and expressive GS representation allows us to achi
eve fast and compact geometry reconstruction photorealistic novel view synthesis
 and effective physically-based rendering. We demonstrate the superiority of our
 method over baseline methods through qualitative and quantitative evaluations o
n various challenging scenes.
****

SynFog: A Photo-realistic Synthetic Fog Dataset based on End-to-end Imaging Simulation for Advancing Real-World Defogging in Autonomous Driving
Yiming Xie, Henglu Wei, Zhenyi Liu, Xiaoyu Wang, Xiangyang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21763-21772
To advance research in learning-based defogging algorithms various synthetic fog datasets have been developed. However exsiting datasets created using the Atmospheric Scattering Model (ASM) or real-time rendering engines often struggle to produce photo-realistic foggy images that accurately mimic the actual imaging process. This limitation hinders the effective generalization of models from synthetic to real data. In this paper we introduce an end-to-end simulation pipeline designed to generate photo-realistic foggy images. This pipeline comprehensively considers the entire physically-based foggy scene imaging process closely aligning with real-world image capture methods. Based on this pipeline we present a new synthetic fog dataset named SynFog which features both sky light and active lighting conditions as well as three levels of fog density. Experimental results demonstrate that models trained on SynFog exhibit superior performance in visual perception and detection accuracy compared to others when applied to real-world foggy images.
****
Video Harmonization with Triplet Spatio-Temporal Variation Patterns
Zonghui Guo, Xinyu Han, Jie Zhang, Shiguang Shan, Haiyong Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19177-19186
Video harmonization is an important and challenging task that aims to obtain visually realistic composite videos by automatically adjusting the foreground's appearance to harmonize with the background. Inspired by the short-term and long-term gradual adjustment process of manual harmonization we present a Video Triplet Transformer framework to model three spatio-temporal variation patterns within videos i.e. short-term spatial as well as long-term global and dynamic for video-to-video tasks like video harmonization. Specifically for short-term harmonization we adjust foreground appearance to consist with background in spatial dimension based on the neighbor frames; for long-term harmonization we not only explore global appearance variations to enhance temporal consistency but also alleviate motion offset constraints to align similar contextual appearances dynamically. Extensive experiments and ablation studies demonstrate the effectiveness of our method achieving state-of-the-art performance in video harmonization video enhancement and video demoireing tasks. We also propose a temporal consistency metric to better evaluate the harmonized videos. Code is available at https://github.com/zhenglab/VideoTripletTransformer.
****
TRINS: Towards Multimodal Language Models that Can Read
Ruiyi Zhang, Yanzhe Zhang, Jian Chen, Yufan Zhou, Jiuxiang Gu, Changyou Chen, Tong Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22584-22594
Large multimodal language models have shown remarkable proficiency in understanding and editing images. However a majority of these visually-tuned models struggle to comprehend the textual content embedded in images primarily due to the limitation of training data. In this work we introduce TRINS: a Text-Rich image1 INStruction dataset with the objective of enhancing the reading ability of the multimodal large language model. TRINS is built upon LAION 2 using hybrid data annotation strategies that include machine-assisted and human-assisted annotation process. It contains 39153 text-rich images captions and 102437 questions. Specifically we show that the number of words per annotation in TRINS is significantly longer than that of related datasets providing new challenges. Furthermore we introduce a simple and effective architecture called a Language-Vision Reading Assistant (LaRA) which is good at understanding textual content within images. LaRA outperforms existing state-of-the-art multimodal large language models on the TRINS dataset as well as other classical benchmarks. Lastly we conducted a comprehensive evaluation with TRINS on various text-rich image understanding and gener

ation tasks demonstrating its effectiveness.
****
Self-Supervised Representation Learning from Arbitrary Scenarios
Zhaowen Li, Yousong Zhu, Zhiyang Chen, Zongxin Gao, Rui Zhao, Chaoyang Zhao, Ming Tang, Jinqiao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22967-22977
Current self-supervised methods can primarily be categorized into contrastive learning and masked image modeling. Extensive studies have demonstrated that combining these two approaches can achieve state-of-the-art performance. However these methods essentially reinforce the global consistency of contrastive learning without taking into account the conflicts between these two approaches which hinders their generalizability to arbitrary scenarios. In this paper we theoretically prove that MAE serves as a patch-level contrastive learning where each patch within an image is considered as a distinct category. This presents a significant conflict with global-level contrastive learning which treats all patches in an image as an identical category. To address this conflict this work abandons the non-generalizable global-level constraints and proposes explicit patch-level contrastive learning as a solution. Specifically this work employs the encoder of MAE to generate dual-branch features which then perform patch-level learning through a decoder. In contrast to global-level data augmentation in contrastive learning our approach leverages patch-level feature augmentation to mitigate interference from global-level learning. Consequently our approach can learn heterogeneous representations from a single image while avoiding the conflicts encountered by previous methods. Massive experiments affirm the potential of our method for learning from arbitrary scenarios.
****
Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions
Oindrila Saha, Grant Van Horn, Subhransu Maji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17542-17552
The zero-shot performance of existing vision-language models (VLMs) such as CLIP is limited by the availability of large-scale aligned image and text datasets in specific domains. In this work we leverage two complementary sources of information -- descriptions of categories generated by large language models (LLMs) and abundant fine-grained image classification datasets -- to improve the zero-shot classification performance of VLMs across fine-grained domains. On the technical side we develop methods to train VLMs with this "bag-level" image-text supervision. We find that simply using these attributes at test-time does not improve performance but our training strategy for example on the iNaturalist dataset leads to an average improvement of 4-5% in zero-shot classification accuracy for novel categories of birds and flowers. Similar improvements are observed in domains where a subset of the categories was used to fine-tune the model. By prompting LLMs in various ways we generate descriptions that capture visual appearance habitat and geographic regions and pair them with existing attributes such as the taxonomic structure of the categories. We systematically evaluate their ability to improve zero-shot categorization in natural domains. Our findings suggest that geographic priors can be just as effective and are complementary to visual appearance. Our method also outperforms prior work on prompt-based tuning of VLMs. We release the benchmark consisting of 14 datasets at https://github.com/cvl-umass/AdaptCLIPZS which will contribute to future research in zero-shot recognition.
****
Living Scenes: Multi-object Relocalization and Reconstruction in Changing 3D Environments
Liyuan Zhu, Shengyu Huang, Konrad Schindler, Iro Armeni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28014-28024
Research into dynamic 3D scene understanding has primarily focused on short-term change tracking from dense observations while little attention has been paid to long-term changes with sparse observations. We address this gap with MoRE a novel approach for multi-object relocalization and reconstruction in evolving envir

onments. We view these environments as Living Scenes and consider the problem of transforming scans taken at different points in time into a 3D reconstruction of the object instances whose accuracy and completeness increase over time. At the core of our method lies an SE(3) equivariant representation in a single encoder-decoder network trained on synthetic data. This representation enables us to seamlessly tackle instance matching registration and reconstruction. We also introduce a joint optimization algorithm that facilitates the accumulation of point clouds originating from the same instance across multiple scans taken at different points in time. We validate our method on synthetic and real-world data and demonstrate state-of-the-art performance in both end-to-end performance and individual subtasks.
****
CricaVPR: Cross-image Correlation-aware Representation Learning for Visual Place Recognition
Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, Chun Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16772-16782
Over the past decade most methods in visual place recognition (VPR) have used neural networks to produce feature representations. These networks typically produce a global representation of a place image using only this image itself and neglect the cross-image variations (e.g. viewpoint and illumination) which limits their robustness in challenging scenes. In this paper we propose a robust global representation method with cross-image correlation awareness for VPR named CricaVPR. Our method uses the attention mechanism to correlate multiple images within a batch. These images can be taken in the same place with different conditions or viewpoints or even captured from different places. Therefore our method can utilize the cross-image variations as a cue to guide the representation learning which ensures more robust features are produced. To further facilitate the robustness we propose a multi-scale convolution-enhanced adaptation method to adapt pre-trained visual foundation models to the VPR task which introduces the multi-scale local information to further enhance the cross-image correlation-aware representation. Experimental results show that our method outperforms state-of-the-art methods by a large margin with significantly less training time. The code is released at https://github.com/Lu-Feng/CricaVPR.
****
ECLIPSE: A Resource-Efficient Text-to-Image Prior for Image Generations
Maitreya Patel, Changhoon Kim, Sheng Cheng, Chitta Baral, Yezhou Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9069-9078
Text-to-image (T2I) diffusion models notably the unCLIP models (e.g. DALL-E-2) achieve state-of-the-art (SOTA) performance on various compositional T2I benchmarks at the cost of significant computational resources. The unCLIP stack comprises T2I prior and diffusion image decoder. The T2I prior model alone adds a billion parameters compared to the Latent Diffusion Models which increases the computational and high-quality data requirements. We introduce ECLIPSE a novel contrastive learning method that is both parameter and data-efficient. ECLIPSE leverages pre-trained vision-language models (e.g. CLIP) to distill the knowledge into the prior model. We demonstrate that the ECLIPSE trained prior with only 3.3% of the parameters and trained on a mere 2.8% of the data surpasses the baseline T2I priors with an average of 71.6% preference score under resource-limited setting. It also attains performance on par with SOTA big models achieving an average of 63.36% preference score in terms of the ability to follow the text compositions. Extensive experiments on two unCLIP diffusion image decoders Karlo and Kandinsky affirm that ECLIPSE priors consistently deliver high performance while significantly reducing resource dependency. Project page: https://eclipse-t2i.vercel.app/
****
Adaptive Bidirectional Displacement for Semi-Supervised Medical Image Segmentation
Hanyang Chi, Jian Pang, Bingfeng Zhang, Weifeng Liu; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4070-4080
Consistency learning is a central strategy to tackle unlabeled data in semi-supervised medical image segmentation (SSMIS) which enforces the model to produce consistent predictions under the perturbation. However most current approaches solely focus on utilizing a specific single perturbation which can only cope with limited cases while employing multiple perturbations simultaneously is hard to guarantee the quality of consistency learning. In this paper we propose an Adaptive Bidirectional Displacement (ABD) approach to solve the above challenge. Specifically we first design a bidirectional patch displacement based on reliable prediction confidence for unlabeled data to generate new samples which can effectively suppress uncontrollable regions and still retain the influence of input perturbations. Meanwhile to enforce the model to learn the potentially uncontrollable content a bidirectional displacement operation with inverse confidence is proposed for the labeled images which generates samples with more unreliable information to facilitate model learning. Extensive experiments show that ABD achieves new state-of-the-art performances for SSMIS significantly improving different baselines. Source code is available at https://github.com/chy-upc/ABD.
****
Accurate Training Data for Occupancy Map Prediction in Automated Driving Using Evidence Theory
Jonas Kälble, Sascha Wirges, Maxim Tatarchenko, Eddy Ilg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5281-5290
Automated driving fundamentally requires knowledge about the surrounding geometry of the scene. Modern approaches use only captured images to predict occupancy maps that represent the geometry. Training these approaches requires accurate data that may be acquired with the help of LiDAR scanners. We show that the techniques used for current benchmarks and training datasets to convert LiDAR scans into occupancy grid maps yield very low quality and subsequently present a novel approach using evidence theory that yields more accurate reconstructions. We demonstrate that these are superior by a large margin both qualitatively and quantitatively and that we additionally obtain meaningful uncertainty estimates. When converting the occupancy maps back to depth estimates and comparing them with the raw LiDAR measurements our method yields a MAE improvement of 30% to 52% on nuScenes and 53% on Waymo over other occupancy ground-truth data. Finally we use the improved occupancy maps to train a state-of-the-art occupancy prediction method and demonstrate that it improves the MAE by 25% on nuScenes.
****
DiffusionLight: Light Probes for Free by Painting a Chrome Ball
Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, Supasorn Suwajanakorn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 98-108
We present a simple yet effective technique to estimate lighting in a single input image. Current techniques rely heavily on HDR panorama datasets to train neural networks to regress an input with limited field-of-view to a full environment map. However these approaches often struggle with real-world uncontrolled settings due to the limited diversity and size of their datasets. To address this problem we leverage diffusion models trained on billions of standard images to render a chrome ball into the input image. Despite its simplicity this task remains challenging: the diffusion models often insert incorrect or inconsistent objects and cannot readily generate chrome balls in HDR format. Our research uncovers a surprising relationship between the appearance of chrome balls and the initial diffusion noise map which we utilize to consistently generate high-quality chrome balls. We further fine-tune an LDR diffusion model (Stable Diffusion XL) with LoRA enabling it to perform exposure bracketing for HDR light estimation. Our method produces convincing light estimates across diverse settings and demonstrates superior generalization to in-the-wild scenarios.
****

Instance-level Expert Knowledge and Aggregate Discriminative Attention for Radio logy Report Generation
Shenshen Bu, Taiji Li, Yuedong Yang, Zhiming Dai; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14194-1420 4
Automatic radiology report generation can provide substantial advantages to clin ical physicians by effectively reducing their workload and improving efficiency. Despite the promising potential of current methods challenges persist in effect ively extracting and preventing degradation of prominent features as well as enh ancing attention on pivotal regions. In this paper we propose an Instance-level Expert Knowledge and Aggregate Discriminative Attention framework (EKAGen) for r adiology report generation. We convert expert reports into an embedding space an d generate comprehensive representations for each disease which serve as Prelimi nary Knowledge Support (PKS). To prevent feature disruption we select the repres entations in the embedding space with the smallest distances to PKS as Rectified Knowledge Support (RKS). Then EKAGen diagnoses the diseases and retrieves knowl edge from RKS creating Instance-level Expert Knowledge (IEK) for each query imag e boosting generation. Additionally we introduce Aggregate Discriminative Attent ion Map (ADM) which uses weak supervision to create maps of discriminative regio ns that highlight pivotal regions. For training we propose a Global Information Self-Distillation (GID) strategy using an iteratively optimized model to distill global knowledge into EKAGen. Extensive experiments and analyses on IU X-Ray an d MIMIC-CXR datasets demonstrate that EKAGen outperforms previous state-of-the-a rt methods.
****
Task-Adaptive Saliency Guidance for Exemplar-free Class Incremental Learning
Xialei Liu, Jiang-Tian Zhai, Andrew D. Bagdanov, Ke Li, Ming-Ming Cheng; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23954-23963
Exemplar-free Class Incremental Learning (EFCIL) aims to sequentially learn task s with access only to data from the current one. EFCIL is of interest because it mitigates concerns about privacy and long-term storage of data while at the sam e time alleviating the problem of catastrophic forgetting in incremental learnin g. In this work we introduce task-adaptive saliency for EFCIL and propose a new framework which we call Task-Adaptive Saliency Supervision (TASS) for mitigating the negative effects of saliency drift between different tasks. We first apply boundary-guided saliency to maintain task adaptivity and plasticity on model att ention. Besides we introduce task-agnostic low-level signals as auxiliary superv ision to increase the stability of model attention. Finally we introduce a modul e for injecting and recovering saliency noise to increase the robustness of sali ency preservation. Our experiments demonstrate that our method can better preser ve saliency maps across tasks and achieve state-of-the-art results on the CIFAR-100 Tiny-ImageNet and ImageNet-Subset EFCIL benchmarks. Code is available at htt ps://github.com/scok30/tass.
****
Rethinking the Spatial Inconsistency in Classifier-Free Diffusion Guidance
Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, Yu Liu; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9370-9379
Classifier-Free Guidance (CFG) has been widely used in text-to-image diffusion m odels where the CFG scale is introduced to control the strength of text guidance on the whole image space. However we argue that a global CFG scale results in s patial inconsistency on varying semantic strengths and suboptimal image quality. To address this problem we present a novel approach Semantic-aware Classifier-F ree Guidance (S-CFG) to customize the guidance degrees for different semantic un its in text-to-image diffusion models. Specifically we first design a training-f ree semantic segmentation method to partition the latent image into relatively i ndependent semantic regions at each denoising step. In particular the cross-atte ntion map in the denoising U-net backbone is renormalized for assigning each pat ch to the corresponding token while the self-attention map is used to complete t

he semantic regions. Then to balance the amplification of diverse semantic units we adaptively adjust the CFG scales across different semantic regions to rescale the text guidance degrees into a uniform level. Finally extensive experiments demonstrate the superiority of S-CFG over the original CFG strategy on various text-to-image diffusion models without requiring any extra training cost. our codes are available at https://github.com/SmilesDZgk/S-CFG.
****
Language-driven All-in-one Adverse Weather Removal
Hao Yang, Liyuan Pan, Yan Yang, Wei Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24902-24912
All-in-one (AiO) frameworks restore various adverse weather degradations with a single set of networks jointly. To handle various weather conditions an AiO framework is expected to adaptively learn weather-specific knowledge for different degradations and shared knowledge for common patterns. However existing method: 1) rely on extra supervision signals which are usually unknown in real-world applications; 2) employ fixed network structures which restrict the diversity of weather-specific knowledge. In this paper we propose a Language-driven Restoration framework (LDR) to alleviate the aforementioned issues. First we leverage the power of pre-trained vision-language (PVL) models to enrich the diversity of weather-specific knowledge by reasoning about the occurrence type and severity of degradation generating description-based degradation priors. Then with the guidance of degradation prior we sparsely select restoration experts from a candidate list dynamically based on a Mixture-of-Experts (MoE) structure. This enables us to adaptively learn the weather-specific and shared knowledge to handle various weather conditions (e.g. unknown or mixed weather). Experiments on extensive restoration scenarios show our superior performance (see Fig. 1). The source code will be made available.
****
Each Test Image Deserves A Specific Prompt: Continual Test-Time Adaptation for 2D Medical Image Segmentation
Ziyang Chen, Yongsheng Pan, Yiwen Ye, Mengkang Lu, Yong Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11184-11193
Distribution shift widely exists in medical images acquired from different medical centres and poses a significant obstacle to deploying the pre-trained semantic segmentation model in real-world applications. Test-time adaptation has proven its effectiveness in tackling the cross-domain distribution shift during inference. However most existing methods achieve adaptation by updating the pre-trained models rendering them susceptible to error accumulation and catastrophic forgetting when encountering a series of distribution shifts (i.e. under the continual test-time adaptation setup). To overcome these challenges caused by updating the models in this paper we freeze the pre-trained model and propose the Visual Prompt-based Test-Time Adaptation (VPTTA) method to train a specific prompt for each test image to align the statistics in the batch normalization layers. Specifically we present the low-frequency prompt which is lightweight with only a few parameters and can be effectively trained in a single iteration. To enhance prompt initialization we equip VPTTA with a memory bank to benefit the current prompt from previous ones. Additionally we design a warm-up mechanism which mixes source and target statistics to construct warm-up statistics thereby facilitating the training process. Extensive experiments demonstrate the superiority of our VPTTA over other state-of-the-art methods on two medical image segmentation benchmark tasks. The code and weights of pre-trained source models are available at https://github.com/Chen-Ziyang/VPTTA.
****
KTPFormer: Kinematics and Trajectory Prior Knowledge-Enhanced Transformer for 3D Human Pose Estimation
Jihua Peng, Yanghong Zhou, P. Y. Mok; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1123-1132
This paper presents a novel Kinematics and Trajectory Prior Knowledge-Enhanced Transformer (KTPFormer) which overcomes the weakness in existing transformer-base

d methods for 3D human pose estimation that the derivation of Q K V vectors in their self-attention mechanisms are all based on simple linear mapping. We propose two prior attention modules namely Kinematics Prior Attention (KPA) and Trajectory Prior Attention (TPA) to take advantage of the known anatomical structure of the human body and motion trajectory information to facilitate effective learning of global dependencies and features in the multi-head self-attention. KPA models kinematic relationships in the human body by constructing a topology of kinematics while TPA builds a trajectory topology to learn the information of joint motion trajectory across frames. Yielding Q K V vectors with prior knowledge the two modules enable KTPFormer to model both spatial and temporal correlations simultaneously. Extensive experiments on three benchmarks (Human3.6M MPI-INF-3DHP and HumanEva) show that KTPFormer achieves superior performance in comparison to state-of-the-art methods. More importantly our KPA and TPA modules have lightweight plug-and-play designs and can be integrated into various transformer-based networks (i.e. diffusion-based) to improve the performance with only a very small increase in the computational overhead. The code is available at: https://github.com/JihuaPeng/KTPFormer.
****
MAPLM: A Real-World Large-Scale Vision-Language Benchmark for Map and Traffic Scene Understanding
Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M. Rehg, Chao Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21819-21830
Vision-language generative AI has demonstrated remarkable promise for empowering cross-modal scene understanding of autonomous driving and high-definition (HD) map systems. However current benchmark datasets lack multi-modal point cloud image and language data pairs. Recent approaches utilize visual instruction learning and cross-modal prompt engineering to expand vision-language models into this domain. In this paper we propose a new vision-language benchmark that can be used to finetune traffic and HD map domain-specific foundation models. Specifically we annotate and leverage large-scale broad-coverage traffic and map data extracted from huge HD map annotations and use CLIP and LLaMA-2 / Vicuna to finetune a baseline model with instruction-following data. Our experimental results across various algorithms reveal that while visual instruction-tuning large language models (LLMs) can effectively learn meaningful representations from MAPLM-QA there remains significant room for further advancements. To facilitate applying LLMs and multi-modal data into self-driving research we will release our visual-language QA data and the baseline models at GitHub.com/LLVM-AD/MAPLM.
****
EgoExoLearn: A Dataset for Bridging Asynchronous Ego- and Exo-centric View of Procedural Activities in Real World
Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22072-22086
Being able to map the activities of others into one's own point of view is one fundamental human skill even from a very early age. Taking a step toward understanding this human ability we introduce EgoExoLearn a large-scale dataset that emulates the human demonstration following process in which individuals record egocentric videos as they execute tasks guided by demonstration videos. Focusing on the potential applications of daily assistance and professional support EgoExoLearn contains egocentric and demonstration video data spanning 120 hours captured in daily life scenarios and specialized laboratories. Along with the videos we record high-quality gaze data and provide detailed multimodal annotations formulating a playground for modeling the human ability to bridge asynchronous procedural actions from different viewpoints. To this end we present benchmarks such as cross-view association cross-view action planning and cross-view referenced skill assessment along with detailed analysis. We expect EgoExoLearn can serve as an important resource for bridging the actions across views thus paving the way for creating AI agents capable of seamlessly learning by observing humans in the

real world. The dataset and benchmark codes are available at https://github.com/OpenGVLab/EgoExoLearn.
****
Differentiable Micro-Mesh Construction
Yishun Dou, Zhong Zheng, Qiaoqiao Jin, Rui Shi, Yuhan Li, Bingbing Ni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4294-4303
Micro-mesh (u-mesh) is a new graphics primitive for compact representation of extreme geometry consisting of a low-polygon base mesh enriched by per micro-vertex displacement. A new generation of GPUs supports this structure with hardware evolution on u-mesh ray tracing achieving real-time rendering in pixel level geometric details. In this article we present a differentiable framework to convert standard meshes into this efficient format offering a holistic scheme in contrast to the previous stage-based methods. In our construction context a u-mesh is defined where each base triangle is a parametric primitive which is then reparameterized with Laplacian operators for efficient geometry optimization. Our framework offers numerous advantages for high-quality u-mesh production: (i) end-to-end geometry optimization and displacement baking; (ii) enabling the differentiation of renderings with respect to umesh for faithful reprojectability; (iii) high scalability for integrating useful features for u-mesh production and rendering such as minimizing shell volume maintaining the isotropy of the base mesh and visual-guided adaptive level of detail. Extensive experiments on u-mesh construction for a large set of high-resolution meshes demonstrate the superior quality achieved by the proposed scheme.
****
Improved Implicit Neural Representation with Fourier Reparameterized Training
Kexuan Shi, Xingyu Zhou, Shuhang Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25985-25994
Implicit Neural Representation (INR) as a mighty representation paradigm has achieved success in various computer vision tasks recently. Due to the low-frequency bias issue of vanilla multi-layer perceptron (MLP) existing methods have investigated advanced techniques such as positional encoding and periodic activation function to improve the accuracy of INR. In this paper we connect the network training bias with the reparameterization technique and theoretically prove that weight reparameterization could provide us a chance to alleviate the spectral bias of MLP. Based on our theoretical analysis we propose a Fourier reparameterization method which learns coefficient matrix of fixed Fourier bases to compose the weights of MLP. We evaluate the proposed Fourier reparameterization method on different INR tasks with various MLP architectures including vanilla MLP MLP with positional encoding and MLP with advanced activation function etc. The superiority approximation results on different MLP architectures clearly validate the advantage of our proposed method. Armed with our Fourier reparameterization method better INR with more textures and less artifacts can be learned from the training data. The codes are available at https://github.com/LabShuHangGU/FR-INR.
****
SNED: Superposition Network Architecture Search for Efficient Video Diffusion Model
Zhengang Li, Yan Kang, Yuchen Liu, Difan Liu, Tobias Hinz, Feng Liu, Yanzhi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8661-8670
While AI-generated content has garnered significant attention achieving photo-realistic video synthesis remains a formidable challenge. Despite the promising advances in diffusion models for video generation quality the complex model architecture and substantial computational demands for both training and inference create a significant gap between these models and real-world applications. This paper presents SNED a superposition network architecture search method for efficient video diffusion model. Our method employs a supernet training paradigm that targets various model cost and resolution options using a weight-sharing method. Moreover we propose the supernet training sampling warm-up for fast training optimization. To showcase the flexibility of our method we conduct experiments invol

ving both pixel-space and latent-space video diffusion models. The results demon strate that our framework consistently produces comparable results across differ ent model options with high efficiency. According to the experiment for the pixe l-space video diffusion model we can achieve consistent video generation results simultaneously across 64 x 64 to 256 x 256 resolutions with a large range of mo del sizes from 640M to 1.6B number of parameters for pixel-space video diffusion models.
****
Groupwise Query Specialization and Quality-Aware Multi-Assignment for Transforme r-based Visual Relationship Detection
Jongha Kim, Jihwan Park, Jinyoung Park, Jinyoung Kim, Sehyung Kim, Hyunwoo J. Ki m; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2024, pp. 28160-28169
Visual Relationship Detection (VRD) has seen significant advancements with Trans former-based architectures recently. However we identify two key limitations in a conventional label assignment for training Transformer-based VRD models which is a process of mapping a ground-truth (GT) to a prediction. Under the conventio nal assignment an 'unspecialized' query is trained since a query is expected to detect every relation which makes it difficult for a query to specialize in spec ific relations. Furthermore a query is also insufficiently trained since a GT is assigned only to a single prediction therefore near-correct or even correct pre dictions are suppressed by being assigned 'no relation' as a GT. To address thes e issues we propose Groupwise Query Specialization and Quality-Aware Multi-Assig nment (SpeaQ). Groupwise Query Specialization trains a 'specialized' query by di viding queries and relations into disjoint groups and directing a query in a spe cific query group solely toward relations in the corresponding relation group. Q uality-Aware Multi-Assignment further facilitates the training by assigning a GT to multiple predictions that are significantly close to a GT in terms of a subj ect an object and the relation in between. Experimental results and analyses sho w that SpeaQ effectively trains 'specialized' queries which better utilize the c apacity of a model resulting in consistent performance gains with 'zero' additio nal inference cost across multiple VRD models and benchmarks. Code is available at https://github.com/mlvlab/SpeaQ.
****
LeftRefill: Filling Right Canvas based on Left Reference through Generalized Tex t-to-Image Diffusion Model
Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, Yanwei Fu; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7705-7715
This paper introduces LeftRefill an innovative approach to efficiently harness l arge Text-to-Image (T2I) diffusion models for reference-guided image synthesis. As the name implies LeftRefill horizontally stitches reference and target views together as a whole input. The reference image occupies the left side while the target canvas is positioned on the right. Then LeftRefill paints the right-side target canvas based on the left-side reference and specific task instructions. S uch a task formulation shares some similarities with contextual inpainting akin to the actions of a human painter. This novel formulation efficiently learns bot h structural and textured correspondence between reference and target without ot her image encoders or adapters. We inject task and view information through cros s-attention modules in T2I models and further exhibit multi-view reference abili ty via the re-arranged self-attention modules. These enable LeftRefill to perfor m consistent generation as a generalized model without requiring test-time fine-tuning or model modifications. Thus LeftRefill can be seen as a simple yet unifi ed framework to address reference-guided synthesis. As an exemplar we leverage L eftRefill to address two different challenges: reference-guided inpainting and n ovel view synthesis based on the pre-trained StableDiffusion. Codes and models a re released at https://github.com/ewrfcas/LeftRefill.
****
Personalized Residuals for Concept-Driven Text-to-Image Generation
Cusuh Ham, Matthew Fisher, James Hays, Nicholas Kolkin, Yuchen Liu, Richard Zhan

g, Tobias Hinz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2024, pp. 8186-8195
We present personalized residuals and localized attention-guided sampling for ef
ficient concept-driven generation using text-to-image diffusion models. Our meth
od first represents concepts by freezing the weights of a pretrained text-condit
ioned diffusion model and learning low-rank residuals for a small subset of the
model's layers. The residual-based approach then directly enables application of
 our proposed sampling technique which applies the learned residuals only in are
as where the concept is localized via cross-attention and applies the original d
iffusion weights in all other regions. Localized sampling therefore combines the
 learned identity of the concept with the existing generative prior of the under
lying diffusion model. We show that personalized residuals effectively capture t
he identity of a concept in 3 minutes on a single GPU without the use of regula
rization images and with fewer parameters than previous models and localized sam
pling allows using the original model as strong prior for large parts of the ima
ge.
****
Condition-Aware Neural Network for Controlled Image Generation
Han Cai, Muyang Li, Qinsheng Zhang, Ming-Yu Liu, Song Han; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7
194-7203
We present Condition-Aware Neural Network (CAN) a new method for adding control
to image generative models. In parallel to prior conditional control methods CAN
 controls the image generation process by dynamically manipulating the weight of
 the neural network. This is achieved by introducing a condition-aware weight ge
neration module that generates conditional weight for convolution/linear layers
based on the input condition. We test CAN on class-conditional image generation
on ImageNet and text-to-image generation on COCO. CAN consistently delivers sign
ificant improvements for diffusion transformer models including DiT and UViT. In
 particular CAN combined with EfficientViT (CaT) achieves 2.78 FID on ImageNet 5
12x512 surpassing DiT-XL/2 while requiring 52x fewer MACs per sampling step.
****
Versatile Navigation Under Partial Observability via Value-guided Diffusion Poli
cy
Gengyu Zhang, Hao Tang, Yan Yan; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2024, pp. 17943-17951
Route planning for navigation under partial observability plays a crucial role i
n modern robotics and autonomous driving. Existing route planning approaches can
 be categorized into two main classes: traditional autoregressive and diffusion-
based methods. The former often fails due to its myopic nature while the latter
either assumes full observability or struggles to adapt to unfamiliar scenarios
due to strong couplings with behavior cloning from experts. To address these def
iciencies we propose a versatile diffusion-based approach for both 2D and 3D rou
te planning under partial observability. Specifically our value-guided diffusion
 policy first generates plans to predict actions across various timesteps provid
ing ample foresight to the planning. It then employs a differentiable planner wi
th state estimations to derive a value function directing the agent's exploratio
n and goal-seeking behaviors without seeking experts while explicitly addressing
 partial observability. During inference our policy is further enhanced by a bes
t-plan-selection strategy substantially boosting the planning success rate. More
over we propose projecting point clouds derived from RGB-D inputs onto 2D grid-b
ased bird-eye-view maps via semantic segmentation generalizing to 3D environment
s. This simple yet effective adaption enables zero-shot transfer from 2D-trained
 policy to 3D cutting across the laborious training for 3D policy and thus certi
fying our versatility. Experimental results demonstrate our superior performance
 particularly in navigating situations beyond expert demonstrations surpassing s
tate-of-the-art autoregressive and diffusion-based baselines for both 2D and 3D
scenarios.
****
All in One Framework for Multimodal Re-identification in the Wild

He Li, Mang Ye, Ming Zhang, Bo Du; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2024, pp. 17459-17469
In Re-identification (ReID) recent advancements yield noteworthy progress in bot
h unimodal and cross-modal retrieval tasks. However the challenge persists in de
veloping a unified framework that could effectively handle varying multimodal da
ta including RGB infrared sketches and textual information. Additionally the eme
rgence of large-scale models shows promising performance in various vision tasks
 but the foundation model in ReID is still blank. In response to these challenge
s a novel multimodal learning paradigm for ReID is introduced referred to as All
-in-One (AIO) which harnesses a frozen pre-trained big model as an encoder enabl
ing effective multimodal retrieval without additional fine-tuning. The diverse m
ultimodal data in AIO are seamlessly tokenized into a unified space allowing the
 modality-shared frozen encoder to extract identity-consistent features comprehe
nsively across all modalities. Furthermore a meticulously crafted ensemble of cr
oss-modality heads is designed to guide the learning trajectory. AIO is the firs
t framework to perform all-in-one ReID encompassing four commonly used modalitie
s. Experiments on cross-modal and multimodal ReID reveal that AIO not only adept
ly handles various modal data but also excels in challenging contexts showcasing
 exceptional performance in zero-shot and domain generalization scenarios. Code
will be available at: https://github.com/lihe404/AIO.
****
Looking 3D: Anomaly Detection with 2D-3D Alignment
Ankan Bhunia, Changjian Li, Hakan Bilen; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17263-17272
Automatic anomaly detection based on visual cues holds practical significance in
 various domains such as manufacturing and product quality assessment. This pape
r introduces a new conditional anomaly detection problem which involves identify
ing anomalies in a query image by comparing it to a reference shape. To address
this challenge we have created a large dataset BrokenChairs-180K consisting of a
round 180K images with diverse anomalies geometries and textures paired with 814
3 reference 3D shapes. To tackle this task we have proposed a novel transformer-
based approach that explicitly learns the correspondence between the query image
 and reference 3D shape via feature alignment and leverages a customized attenti
on mechanism for anomaly detection. Our approach has been rigorously evaluated t
hrough comprehensive experiments serving as a benchmark for future research in t
his domain.
****
Purified and Unified Steganographic Network
Guobiao Li, Sheng Li, Zicong Luo, Zhenxing Qian, Xinpeng Zhang; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 27569-27578
Steganography is the art of hiding secret data into the cover media for covert c
ommunication. In recent years more and more deep neural network (DNN)-based steg
anographic schemes are proposed to train steganographic networks for secret embe
dding and recovery which are shown to be promising. Compared with the handcrafte
d steganographic tools steganographic networks tend to be large in size. It rais
es concerns on how to imperceptibly and effectively transmit these networks to t
he sender and receiver to facilitate the covert communication. To address this i
ssue we propose in this paper a Purified and Unified Steganographic Network (PUS
Net). It performs an ordinary machine learning task in a purified network which
could be triggered into steganographic networks for secret embedding or recovery
 using different keys. We formulate the construction of the PUSNet into a sparse
 weight filling problem to flexibly switch between the purified and steganograph
ic networks. We further instantiate our PUSNet as an image denoising network wit
h two steganographic networks concealed for secret image embedding and recovery.
 Comprehensive experiments demonstrate that our PUSNet achieves good performance
 on secret image embedding secret image recovery and image denoising in a single
 architecture. It is also shown to be capable of imperceptibly carrying the steg
anographic networks in a purified network. Steganography is the art of hiding se
cret data into the cover media for covert communication. In recent years more an

d more deep neural network (DNN)-based steganographic schemes are proposed to train steganographic networks for secret embedding and recovery which are shown to be promising. Compared with the handcrafted steganographic tools steganographic networks tend to be large in size. It raises concerns on how to imperceptibly and effectively transmit these networks to the sender and receiver to facilitate the covert communication. To address this issue we propose in this paper a Purified and Unified Steganographic Network (PUSNet). It performs an ordinary machine learning task in a purified network which could be triggered into steganographic networks for secret embedding or recovery using different keys. We formulate the construction of the PUSNet into a sparse weight filling problem to flexibly switch between the purified and steganographic networks. We further instantiate our PUSNet as an image denoising network with two steganographic networks concealed for secret image embedding and recovery. Comprehensive experiments demonstrate that our PUSNet achieves good performance on secret image embedding secret image recovery and image denoising in a single architecture. It is also shown to be capable of imperceptibly carrying the steganographic networks in a purified network. Code is available at https://github.com/albblgb/PUSNet
****
VS: Reconstructing Clothed 3D Human from Single Image via Vertex Shift
Leyuan Liu, Yuhan Li, Yunqi Gao, Changxin Gao, Yuanyuan Liu, Jingying Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10498-10507
Various applications require high-fidelity and artifact-free 3D human reconstructions. However current implicit function-based methods inevitably produce artifacts while existing deformation methods are difficult to reconstruct high-fidelity humans wearing loose clothing. In this paper we propose a two-stage deformation method named Vertex Shift (VS) for reconstructing clothed 3D humans from single images. Specifically VS first stretches the estimated SMPL-X mesh into a coarse 3D human model using shift fields inferred from normal maps then refines the coarse 3D human model into a detailed 3D human model via a graph convolutional network embedded with implicit-function-learned features. This "stretch-refine" strategy addresses large deformations required for reconstructing loose clothing and delicate deformations for recovering intricate and detailed surfaces achieving high-fidelity reconstructions that faithfully convey the pose clothing and surface details from the input images. The graph convolutional network's ability to exploit neighborhood vertices coupled with the advantages inherited from the deformation methods ensure VS rarely produces artifacts like distortions and non-human shapes and never produces artifacts like holes broken parts and dismembered limbs. As a result VS can reconstruct high-fidelity and artifact-less clothed 3D humans from single images even under scenarios of challenging poses and loose clothing. Experimental results on three benchmarks and two in-the-wild datasets demonstrate that VS significantly outperforms current state-of-the-art methods. The code and models of VS are available for research purposes at https://github.com/starVisionTeam/VS.
****
PARA-Drive: Parallelized Architecture for Real-time Autonomous Driving
Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, Marco Pavone; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15449-15458
Recent works have proposed end-to-end autonomous vehicle (AV) architectures comprised of differentiable modules achieving state-of-the-art driving performance. While they provide advantages over the traditional perception-prediction-planning pipeline (e.g. removing information bottlenecks between components and alleviating integration challenges) they do so using a diverse combination of tasks modules and their interconnectivity. As of yet however there has been no systematic analysis of the necessity of these modules or the impact of their connectivity placement and internal representations on overall driving performance. Addressing this gap our work conducts a comprehensive exploration of the design space of end-to-end modular AV stacks. Our findings culminate in the development of PARA-Drive: a fully parallel end-to-end AV architecture. PARA-Drive not only achieves

state-of-the-art performance in perception prediction and planning but also significantly enhances runtime speed by nearly 3x without compromising on interpretability or safety.
****
TEA: Test-time Energy Adaptation
Yige Yuan, Bingbing Xu, Liang Hou, Fei Sun, Huawei Shen, Xueqi Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23901-23911
Test Time Adaptation (TTA) aims to improve model generalizability when test data diverges from training distribution with the distinct advantage of not requiring access to training data and processes especially valuable in the context of pre-trained models. However current TTA methods fail to address the fundamental issue: covariate shift i.e. the decreased generalizability can be attributed to the model's reliance on the marginal distribution of the training data which may impair model calibration and introduce confirmation bias. To address this we propose a novel energy-based perspective enhancing the model's perception of target data distributions without requiring access to training data or processes. Building on this perspective we introduce Test-time Energy Adaptation (TEA) which transforms the trained classifier into an energy-based model and aligns the model's distribution with the test data's enhancing its ability to perceive test distributions and thus improving overall generalizability. Extensive experiments across multiple tasks benchmarks and architectures demonstrate TEA's superior generalization performance against state-of-the-art methods. Further in-depth analyses reveal that TEA can equip the model with a comprehensive perception of test distribution ultimately paving the way toward improved generalization and calibration. Code is available at https://github.com/yuanyige/tea.
****
NEAT: Distilling 3D Wireframes from Neural Attraction Fields
Nan Xue, Bin Tan, Yuxi Xiao, Liang Dong, Gui-Song Xia, Tianfu Wu, Yujun Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19968-19977
This paper studies the problem of structured 3D recon- struction using wireframes that consist of line segments and junctions focusing on the computation of structured boundary geometries of scenes. Instead of leveraging matching-based solutions from 2D wireframes (or line segments) for 3D wireframe reconstruction as done in prior arts we present NEAT a rendering-distilling formulation using neural fields to represent 3D line segments with 2D observations and bipartite matching for perceiving and dis- tilling of a sparse set of 3D global junctions. The proposed NEAT enjoys the joint optimization of the neural fields and the global junctions from scratch using view-dependent 2D observations without precomputed cross-view feature matching. Comprehensive experiments on the DTU and BlendedMVS datasets demonstrate our NEAT's superiority over state-of-the-art alternatives for 3D wireframe recon- struction. Moreover the distilled 3D global junctions by NEAT are a better initialization than SfM points for the recently-emerged 3D Gaussian Splatting for high-fidelity novel view synthesis using about 20 times fewer initial 3D points. Project page: https://xuenan.net/neat
****
Prompt Augmentation for Self-supervised Text-guided Image Manipulation
Rumeysa Bodur, Binod Bhattarai, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8829-8838
Text-guided image editing finds applications in various creative and practical fields. While recent studies in image generation have advanced the field they often struggle with the dual challenges of coherent image transformation and context preservation. In response our work introduces prompt augmentation a method amplifying a single input prompt into several target prompts strengthening textual context and enabling localised image editing. Specifically we use the augmented prompts to delineate the intended manipulation area. We propose a Contrastive Loss tailored to driving effective image editing by displacing edited areas and drawing preserved regions closer. Acknowledging the continuous nature of image manipulations we further refine our approach by incorporating the similarity concep

t creating a Soft Contrastive Loss. The new losses are incorporated to the diffu
sion model demonstrating improved or competitive image editing results on public
 datasets and generated images over state-of-the-art approaches.
****
Pink: Unveiling the Power of Referential Comprehension for Multi-modal LLMs
Shiyu Xuan, Qingpei Guo, Ming Yang, Shiliang Zhang; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13838-13
848
Multi-modal Large Language Models (MLLMs) have shown remarkable capabilities in
various multi-modal tasks. Nevertheless their performance in fine-grained image
understanding tasks is still limited. To address this issue this paper proposes
a new framework to enhance the fine-grained image understanding abilities of MLL
Ms. Specifically we present a new method for constructing the instruction tuning
 dataset at a low cost by leveraging annotations in existing datasets. A self-co
nsistent bootstrapping method is also introduced to extend existing dense object
 annotations into high-quality referring-expression-bounding-box pairs. These me
thods enable the generation of high-quality instruction data which includes a wi
de range of fundamental abilities essential for fine-grained image perception. M
oreover we argue that the visual encoder should be tuned during instruction tuni
ng to mitigate the gap between full image perception and fine-grained image perc
eption. Experimental results demonstrate the superior performance of our method.
 For instance our model exhibits a 5.2% accuracy improvement over Qwen-VL on GQA
 and surpasses the accuracy of Kosmos-2 by 24.7% on RefCOCO_val. We have also at
tained the top rank on the leaderboard of MMBench. This promising performance is
 achieved by training on only publicly available data making it easily reproduci
ble. The models datasets and codes are publicly available at https://github.com/
SY-Xuan/Pink.
****
LDP: Language-driven Dual-Pixel Image Defocus Deblurring Network
Hao Yang, Liyuan Pan, Yan Yang, Richard Hartley, Miaomiao Liu; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 24078-24087
Recovering sharp images from dual-pixel (DP) pairs with disparity-dependent blur
 is a challenging task. Existing blur map-based deblurring methods have demonstr
ated promising results. In this paper we propose to the best of our knowledge th
e first framework to introduce the contrastive language-image pre-training frame
work (CLIP) to achieve accurate blur map estimation from DP pairs unsupervisedly
. To this end we first carefully design text prompts to enable CLIP to understan
d blur-related geometric prior knowledge from the DP pair. Then we propose a for
mat to input stereo DP pair to the CLIP without any fine-tuning where the CLIP i
s pre-trained on monocular images. Given the estimated blur map we introduce a b
lur-prior attention block a blur-weighting loss and a blur-aware loss to recover
 the all-in-focus image. Our method achieves state-of-the-art performance in ext
ensive experiments (see Fig. 1).
****
MMSum: A Dataset for Multimodal Summarization and Thumbnail Generation of Videos
Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin
, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, Bo Li, Lijuan Wang; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2024, pp. 21909-21921
Multimodal summarization with multimodal output (MSMO) has emerged as a promisin
g research direction. Nonetheless numerous limitations exist within existing pub
lic MSMO datasets including insufficient maintenance data inaccessibility limite
d size and the absence of proper categorization which pose significant challenge
s. To address these challenges and provide a comprehensive dataset for this new
direction we have meticulously curated the MMSum dataset. Our new dataset featur
es (1) Human-validated summaries for both video and textual content providing su
perior human instruction and labels for multimodal learning. (2) Comprehensively
 and meticulously arranged categorization spanning 17 principal categories and 1
70 subcategories to encapsulate a diverse array of real-world scenarios. (3) Ben

chmark tests performed on the proposed dataset to assess various tasks and metho ds including video summarization text summarization and multimodal summarization . To champion accessibility and collaboration we released the MMSum dataset and the data collection tool as fully open-source resources fostering transparency a nd accelerating future developments.
****
HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data
Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, Yueting Zhuang; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2024, pp. 12944-12953
Multi-modal Large Language Models (MLLMs) tuned on machine-generated instruction -following data have demonstrated remarkable performance in various multimodal u nderstanding and generation tasks. However the hallucinations inherent in machin e-generated data which could lead to hallucinatory outputs in MLLMs remain under -explored. This work aims to investigate various hallucinations (i.e. object rel ation attribute hallucinations) and mitigate those hallucinatory toxicities in l arge-scale machine-generated visual instruction datasets. Drawing on the human a bility to identify factual errors we present a novel hallucination detection and elimination framework HalluciDoctor based on the cross-checking paradigm. We us e our framework to identify and eliminate hallucinations in the training data au tomatically. Interestingly HalluciDoctor also indicates that spurious correlatio ns arising from long-tail object co-occurrences contribute to hallucinations. Ba sed on that we execute counterfactual visual instruction expansion to balance da ta distribution thereby enhancing MLLMs' resistance to hallucinations. Comprehen sive experiments on hallucination evaluation benchmarks show that our method suc cessfully mitigates 44.6% hallucinations relatively and maintains competitive pe rformance compared to LLaVA. The data and code for this paper are publicly avail able.
****
Pre-trained Vision and Language Transformers Are Few-Shot Incremental Learners
Keon-Hee Park, Kyungwoo Song, Gyeong-Moon Park; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23881-23890
Few-Shot Class Incremental Learning (FSCIL) is a task that requires a model to l earn new classes incrementally without forgetting when only a few samples for ea ch class are given. FSCIL encounters two significant challenges: catastrophic fo rgetting and overfitting and these challenges have driven prior studies to prima rily rely on shallow models such as ResNet-18. Even though their limited capacit y can mitigate both forgetting and overfitting issues it leads to inadequate kno wledge transfer during few-shot incremental sessions. In this paper we argue tha t large models such as vision and language transformers pre-trained on large dat asets can be excellent few-shot incremental learners. To this end we propose a n ovel FSCIL framework called PriViLege Pre-trained Vision and Language transforme rs with prompting functions and knowledge distillation. Our framework effectivel y addresses the challenges of catastrophic forgetting and overfitting in large m odels through new pre-trained knowledge tuning (PKT) and two losses: entropy-bas ed divergence loss and semantic knowledge distillation loss. Experimental result s show that the proposed PriViLege significantly outperforms the existing state-of-the-art methods with a large margin e.g. +9.38% in CUB200 +20.58% in CIFAR-10 0 and +13.36% in miniImageNet. Our implementation code is available at https://g ithub.com/KHU-AGI/PriViLege.
****
Guess The Unseen: Dynamic 3D Scene Reconstruction from Partial 2D Glimpses
Inhee Lee, Byungjun Kim, Hanbyul Joo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1062-1071
In this paper we present a method to reconstruct the world and multiple dynamic humans in 3D from a monocular video input. As a key idea we represent both the w orld and multiple humans via the recently emerging 3D Gaussian Splatting (3D-GS) representation enabling to conveniently and efficiently compose and render them together. In particular we address the scenarios with severely limited and spar se observations in 3D human reconstruction a common challenge encountered in the

real world. To tackle this challenge we introduce a novel approach to optimize the 3D-GS representation in a canonical space by fusing the sparse cues in the common space where we leverage a pre-trained 2D diffusion model to synthesize unseen views while keeping the consistency with the observed 2D appearances. We demonstrate our method can reconstruct high-quality animatable 3D humans in various challenging examples in the presence of occlusion image crops few-shot and extremely sparse observations. After reconstruction our method is capable of not only rendering the scene in any novel views at arbitrary time instances but also editing the 3D scene by removing individual humans or applying different motions for each human. Through various experiments we demonstrate the quality and efficiency of our methods over alternative existing approaches.
****
C^2RV: Cross-Regional and Cross-View Learning for Sparse-View CBCT Reconstruction
Yiqun Lin, Jiewen Yang, Hualiang Wang, Xinpeng Ding, Wei Zhao, Xiaomeng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11205-11214
Cone beam computed tomography (CBCT) is an important imaging technology widely used in medical scenarios such as diagnosis and preoperative planning. Using fewer projection views to reconstruct CT also known as sparse-view reconstruction can reduce ionizing radiation and further benefit interventional radiology. Compared with sparse-view reconstruction for traditional parallel/fan-beam CT CBCT reconstruction is more challenging due to the increased dimensionality caused by the measurement process based on cone-shaped X-ray beams. As a 2D-to-3D reconstruction problem although implicit neural representations have been introduced to enable efficient training only local features are considered and different views are processed equally in previous works resulting in spatial inconsistency and poor performance on complicated anatomies. To this end we propose C^2RV by leveraging explicit multi-scale volumetric representations to enable cross-regional learning in the 3D space. Additionally the scale-view cross-attention module is introduced to adaptively aggregate multi-scale and multi-view features. Extensive experiments demonstrate that our C^2RV achieves consistent and significant improvement over previous state-of-the-art methods on datasets with diverse anatomy. Code is available at https://github.com/xmed-lab/C2RV-CBCT.
****
HyperDreamBooth: HyperNetworks for Fast Personalization of Text-to-Image Models
Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, Kfir Aberman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6527-6536
Personalization has emerged as a prominent aspect within the field of generative AI enabling the synthesis of individuals in diverse contexts and styles while retaining high-fidelity to their identities. However the process of personalization presents inherent challenges in terms of time and memory requirements. Fine-tuning each personalized model needs considerable GPU time investment and storing a personalized model per subject can be demanding in terms of storage capacity. To overcome these challenges we propose HyperDreamBooth - a hypernetwork capable of efficiently generating a small set of personalized weights from a single image of a person. By composing these weights into the diffusion model coupled with fast finetuning HyperDreamBooth can generate a person's face in various contexts and styles with high subject details while also preserving the model's crucial knowledge of diverse styles and semantic modifications. Our method achieves personalization on faces in roughly 20 seconds 25x faster than DreamBooth and 125x faster than Textual Inversion using as few as one reference image with the same quality and style diversity as DreamBooth. Also our method yields a model that is 10000x smaller than a normal DreamBooth model.
****
Language-guided Image Reflection Separation
Haofeng Zhong, Yuchen Hong, Shuchen Weng, Jinxiu Liang, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24913-24922

This paper studies the problem of language-guided reflection separation which ai
ms at addressing the ill-posed reflection separation problem by introducing lang
uage descriptions to provide layer content. We propose a unified framework to so
lve this problem which leverages the cross-attention mechanism with contrastive
learning strategies to construct the correspondence between language description
s and image layers. A gated network design and a randomized training strategy ar
e employed to tackle the recognizable layer ambiguity. The effectiveness of the
proposed method is validated by the significant performance advantage over exist
ing reflection separation methods on both quantitative and qualitative compariso
ns.
****
HardMo: A Large-Scale Hardcase Dataset for Motion Capture
Jiaqi Liao, Chuanchen Luo, Yinuo Du, Yuxi Wang, Xucheng Yin, Man Zhang, Zhaoxian
g Zhang, Junran Peng; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2024, pp. 1629-1638
Recent years have witnessed rapid progress in monocular human mesh recovery. Des
pite their impressive performance on public benchmarks existing methods are vuln
erable to unusual poses which prevents them from deploying to challenging scenar
ios such as dance and martial arts. This issue is mainly attributed to the domai
n gap induced by the data scarcity in relevant cases. Most existing datasets are
 captured in constrained scenarios and lack samples of such complex movements. F
or this reason we propose a data collection pipeline comprising automatic crawli
ng precise annotation and hardcase mining. Based on this pipeline we establish a
 large dataset in a short time. The dataset named HardMo contains 7M images alon
g with precise annotations covering 15 categories of dance and 14 categories of
martial arts. Empirically we find that the prediction failure in dance and marti
al arts is mainly characterized by the misalignment of hand-wrist and foot-ankle
. To dig deeper into the two hardcases we leverage the proposed automatic pipeli
ne to filter collected data and construct two subsets named HardMo-Hand and Hard
Mo-Foot. Extensive experiments demonstrate the effectiveness of the annotation p
ipeline and the data-driven solution to failure cases. Specifically after being
trained on HardMo HMR an early pioneering method can even outperform the current
 state of the art 4DHumans on our benchmarks.
****
View-Category Interactive Sharing Transformer for Incomplete Multi-View Multi-La
bel Learning
Shilong Ou, Zhe Xue, Yawen Li, Meiyu Liang, Yuanqiang Cai, Junjiang Wu; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2024, pp. 27467-27476
As a problem often encountered in real-world scenarios multi-view multi-label le
arning has attracted considerable research attention. However due to oversights
in data collection and uncertainties in manual annotation real-world data often
suffer from incompleteness. Regrettably most existing multi-view multi-label lea
rning methods sidestep missing views and labels. Furthermore they often neglect
the potential of harnessing complementary information between views and labels t
hus constraining their classification capabilities. To address these challenges
we propose a view-category interactive sharing transformer tailored for incomple
te multi-view multi-label learning. Within this network we incorporate a two-lay
er transformer module to characterize the interplay between views and labels. Ad
ditionally to address view incompleteness a KNN-style missing view generation mo
dule is employed. Finally we introduce a view-category consistency guided embedd
ing enhancement module to align different views and improve the discriminating p
ower of the embeddings. Collectively these modules synergistically integrate to
classify the incomplete multi-view multi-label data effectively. Extensive exper
iments substantiate that our approach outperforms the existing state-of-the-art
methods.
****
The More You See in 2D the More You Perceive in 3D
Xinyang Han, Zelin Gao, Angjoo Kanazawa, Shubham Goel, Yossi Gandelsman; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR

), 2024, pp. 20912-20922
Humans can infer 3D structure from 2D images of an object based on past experien ce and improve their 3D understanding as they see more images. Inspired by this behavior we introduce SAP3D a system for 3D reconstruction and novel view synthe sis from an arbitrary number of unposed images. Given a few unposed images of an object we adapt a pre-trained view-conditioned diffusion model together with th e camera poses of the images via test-time fine-tuning. The adapted diffusion mo del and the obtained camera poses are then utilized as instance-specific priors for 3D reconstruction and novel view synthesis. We show that as the number of in put images increases the performance of our approach improves bridging the gap b etween optimization-based prior-less 3D reconstruction methods and single-image-to-3D diffusion-based methods. We demonstrate our system on real images as well as standard synthetic benchmarks. Our ablation studies confirm that this adaptio n behavior is key for more accurate 3D understanding.
****
GLiDR: Topologically Regularized Graph Generative Network for Sparse LiDAR Point Clouds
Prashant Kumar, Kshitij Madhav Bhat, Vedang Bhupesh Shenvi Nadkarni, Prem Kalra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2024, pp. 15152-15161
Sparse LiDAR point clouds cause severe loss of detail of static structures and r educe the density of static points available for navigation. Reduced density can be detrimental to navigation under several scenarios. We observe that despite h igh sparsity in most cases the global topology of LiDAR outlining the static str uctures can be inferred. We utilize this property to obtain a backbone skeleton of a LiDAR scan in the form of a single connected component that is a proxy to i ts global topology. We utilize the backbone to augment new points along static s tructures to overcome sparsity. Newly introduced points could correspond to exis ting static structures or to static points that were earlier obstructed by dynam ic objects. To the best of our knowledge we are the first to use such a strategy for sparse LiDAR point clouds. Existing solutions close to our approach fail to identify and preserve the global static LiDAR topology and generate sub-optimal points. We propose GLiDR a Graph Generative network that is topologically regul arized using 0-dimensional Persistent Homology (PH) constraints. This enables GL iDR to introduce newer static points along a topologically consistent global sta tic LiDAR backbone. GLiDR generates precise static points using 32x sparser dyna mic scans and performs better than the baselines across three datasets. GLiDR ge nerates a valuable byproduct - an accurate binary segmentation mask of static an d dynamic objects that are helpful for navigation planning and safety in constra ined environments. The newly introduced static points allow GLiDR to outperform LiDAR-based navigation using SLAM in several settings.
****
Separate and Conquer: Decoupling Co-occurrence via Decomposition and Representat ion for Weakly Supervised Semantic Segmentation
Zhiwei Yang, Kexue Fu, Minghong Duan, Linhao Qu, Shuo Wang, Zhijian Song; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2024, pp. 3606-3615
Weakly supervised semantic segmentation (WSSS) with image-level labels aims to a chieve segmentation tasks without dense annotations. However attributed to the f requent coupling of co-occurring objects and the limited supervision from image-level labels the challenging co-occurrence problem is widely present and leads t o false activation of objects in WSSS. In this work we devise a 'Separate and Co nquer' scheme SeCo to tackle this issue from dimensions of image space and featu re space. In the image space we propose to 'separate' the co-occurring objects w ith image decomposition by subdividing images into patches. Importantly we assig n each patch a category tag from Class Activation Maps (CAMs) which spatially he lps remove the co-context bias and guide the subsequent representation. In the f eature space we propose to 'conquer' the false activation by enhancing semantic representation with multi-granularity knowledge contrast. To this end a dual-tea cher-single-student architecture is designed and tag-guided contrast is conducte

d which guarantee the correctness of knowledge and further facilitate the discrepancy among co-contexts. We streamline the multi-staged WSSS pipeline end-to-end and tackle this issue without external supervision. Extensive experiments are conducted validating the efficiency of our method and the superiority over previous single-staged and even multi-staged competitors on PASCAL VOC and MS COCO. Code is available at https://github.com/zwyang6/SeCo.git.
****

BiPer: Binary Neural Networks using a Periodic Function
Edwin Vargas, Claudia V. Correa, Carlos Hinojosa, Henry Arguello; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5684-5693
Quantized neural networks employ reduced precision representations for both weights and activations. This quantization process significantly reduces the memory requirements and computational complexity of the network. Binary Neural Networks (BNNs) are the extreme quantization case representing values with just one bit. Since the sign function is typically used to map real values to binary values smooth approximations are introduced to mimic the gradients during error backpropagation. Thus the mismatch between the forward and backward models corrupts the direction of the gradient causing training inconsistency problems and performance degradation. In contrast to current BNN approaches we propose to employ a binary periodic (BiPer) function during binarization. Specifically we use a square wave for the forward pass to obtain the binary values and employ the trigonometric sine function with the same period of the square wave as a differentiable surrogate during the backward pass. We demonstrate that this approach can control the quantization error by using the frequency of the periodic function and improves network performance. Extensive experiments validate the effectiveness of BiPer in benchmark datasets and network architectures with improvements of up to 1% and 0.69% with respect to state-of-the-art methods in the classification task over CIFAR-10 and ImageNet respectively. Our code is publicly available at https://github.com/edmav4/BiPer.
****

Unifying Automatic and Interactive Matting with Pretrained ViTs
Zixuan Ye, Wenze Liu, He Guo, Yujia Liang, Chaoyi Hong, Hao Lu, Zhiguo Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25585-25594
Automatic and interactive matting largely improve image matting by respectively alleviating the need for auxiliary input and enabling object selection. Due to different settings on whether prompts exist they either suffer from weakness in instance completeness or region details. Also when dealing with different scenarios directly switching between the two matting models introduces inconvenience and higher workload. Therefore we wonder whether we can alleviate the limitations of both settings while achieving unification to facilitate more convenient use. Our key idea is to offer saliency guidance for automatic mode to enable its attention to detailed regions and also refine the instance completeness in interactive mode by replacing the binary mask guidance with a more probabilistic form. With different guidance for each mode we can achieve unification through adaptable guidance defined as saliency information in automatic mode and user cue for interactive one. It is instantiated as candidate feature in our method an automatic switch for class token in pretrained ViTs and average feature of user prompts controlled by the existence of user prompts. Then we use the candidate feature to generate a probabilistic similarity map as the guidance to alleviate the over-reliance on binary mask. Extensive experiments show that our method can adapt well to both automatic and interactive scenarios with more light-weight framework. Code available at https://github.com/coconuthust/SmartMatting.
****

Segment Any Event Streams via Weighted Adaptation of Pivotal Tokens
Zhiwen Chen, Zhiyu Zhu, Yifan Zhang, Junhui Hou, Guangming Shi, Jinjian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3890-3900
In this paper we delve into the nuanced challenge of tailoring the Segment Anyth

ing Models (SAMs) for integration with event data with the overarching objective of attaining robust and universal object segmentation within the event-centric domain. One pivotal issue at the heart of this endeavor is the precise alignment and calibration of embeddings derived from event-centric data such that they ha rmoniously coincide with those originating from RGB imagery. Capitalizing on the vast repositories of datasets with paired events and RGB images our proposition is to harness and extrapolate the profound knowledge encapsulated within the pr e-trained SAM framework. As a cornerstone to achieving this we introduce a multi -scale feature distillation methodology. This methodology rigorously optimizes t he alignment of token embeddings originating from event data with their RGB imag e counterparts thereby preserving and enhancing the robustness of the overall ar chitecture. Considering the distinct significance that token embeddings from int ermediate layers hold for higher-level embeddings our strategy is centered on ac curately calibrating the pivotal token embeddings. This targeted calibration is aimed at effectively managing the discrepancies in high-level embeddings origina ting from both the event and image domains. Extensive experiments on different d atasets demonstrate the effectiveness of the proposed distillation method. Code in https://github.com/happychenpipi/EventSAM.
****

AnyDoor: Zero-shot Object-level Image Customization
Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, Hengshuang Zhao; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6593-6602
This work presents AnyDoor a diffusion-based image generator with the power to t eleport target objects to new scenes at user-specified locations with desired sh apes. Instead of tuning parameters for each object our model is trained only onc e and effortlessly generalizes to diverse object-scene combinations at the infer ence stage. Such a challenging zero-shot setting requires an adequate characteri zation of a certain object. To this end we complement the commonly used identity feature with detail features which are carefully designed to maintain appearanc e details yet allow versatile local variations(e.g. lighting orientation posture etc.) supporting the object in favorably blending with different surroundings. We further propose to borrow knowledge from video datasets where we can observe various forms (i.e. along the time axis) of a single object leading to stronger model generalizability and robustness. Extensive experiments demonstrate the sup eriority of our approach over existing alternatives as well as its great potenti al in real-world applications such as virtual try-on shape editing and object sw apping.
****

Commonsense Prototype for Outdoor Unsupervised 3D Object Detection
Hai Wu, Shijia Zhao, Xun Huang, Chenglu Wen, Xin Li, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14968-14977
The prevalent approaches of unsupervised 3D object detection follow cluster-base d pseudo-label generation and iterative self-training processes. However the cha llenge arises due to the sparsity of LiDAR scans which leads to pseudo-labels wi th erroneous size and position resulting in subpar detection performance. To tac kle this problem this paper introduces a Commonsense Prototype-based Detector te rmed CPD for unsupervised 3D object detection. CPD first constructs Commonsense Prototype (CProto) characterized by high-quality bounding box and dense points b ased on commonsense intuition. Subsequently CPD refines the low-quality pseudo-l abels by leveraging the size prior from CProto. Furthermore CPD enhances the det ection accuracy of sparsely scanned objects by the geometric knowledge from CPro to. CPD outperforms state-of-the-art unsupervised 3D detectors on Waymo Open Dat aset (WOD) PandaSet and KITTI datasets by a large margin. Besides by training CP D on WOD and testing on KITTI CPD attains 90.85% and 81.01% 3D Average Precision on easy and moderate car classes respectively. These achievements position CPD in close proximity to fully supervised detectors highlighting the significance o f our method. The code will be available at https://github.com/hailanyi/CPD.
****

Lookahead Exploration with Neural Radiance Representation for Continuous Vision-Language Navigation
Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Junjie Hu, Ming Jiang, Shuqiang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13753-13762
Vision-and-language navigation (VLN) enables the agent to navigate to a remote location following the natural language instruction in 3D environments. At each navigation step the agent selects from possible candidate locations and then makes the move. For better navigation planning the lookahead exploration strategy aims to effectively evaluate the agent's next action by accurately anticipating the future environment of candidate locations. To this end some existing works predict RGB images for future environments while this strategy suffers from image distortion and high computational cost. To address these issues we propose the pre-trained hierarchical neural radiance representation model (HNR) to produce multi-level semantic features for future environments which are more robust and efficient than pixel-wise RGB reconstruction. Furthermore with the predicted future environmental representations our lookahead VLN model is able to construct the navigable future path tree and select the optimal path via efficient parallel evaluation. Extensive experiments on the VLN-CE datasets confirm the effectiveness of our method.
****
Clustering Propagation for Universal Medical Image Segmentation
Yuhang Ding, Liulei Li, Wenguan Wang, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3357-3369
Prominent solutions for medical image segmentation are typically tailored for automatic or interactive setups posing challenges in facilitating progress achieved in one task to another. This also necessitates separate models for each task duplicating both training time and parameters. To address above issues we introduce S2VNet a universal framework that leverages Slice-to-Volume propagation to unify automatic/interactive segmentation within a single model and one training session. Inspired by clustering-based segmentation techniques S2VNet makes full use of the slice-wise structure of volumetric data by initializing cluster centers from the cluster results of previous slice. This enables knowledge acquired from prior slices to assist in the segmentation of the current slice further efficiently bridging the communication between remote slices using mere 2D networks. Moreover such a framework readily accommodates inter- active segmentation with no architectural change simply by initializing centroids from user inputs. S2VNet distinguishes itself by swift inference speeds and reduced memory consumption compared to prevailing 3D solutions. It can also handle multi-class interactions with each of them serving to initialize different centroids. Experiments on three benchmarks demonstrate S2VNet surpasses task-specified solutions on both automatic/interactive setups.
****
MoPE-CLIP: Structured Pruning for Efficient Vision-Language Models with Module-wise Pruning Error Metric
Haokun Lin, Haoli Bai, Zhili Liu, Lu Hou, Muyi Sun, Linqi Song, Ying Wei, Zhenan Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27370-27380
Vision-language pre-trained models have achieved impressive performance on various downstream tasks. However their large model sizes hinder their utilization on platforms with limited computational resources. We find that directly using smaller pre-trained models and applying magnitude-based pruning on CLIP models leads to inflexibility and inferior performance. Recent efforts for VLP compression either adopt uni-modal compression metrics resulting in limited performance or involve costly mask-search processes with learnable masks. In this paper we first propose the Module-wise Pruning Error (MoPE) metric accurately assessing CLIP module importance by performance decline on cross-modal tasks. Using the MoPE metric we introduce a unified pruning framework applicable to both pre-training and task-specific fine-tuning compression stages. For pre-training MoPE-CLIP effectively leverages knowledge from the teacher model significantly reducing pre-trai

ning costs while maintaining strong zero-shot capabilities. For fine-tuning cons
ecutive pruning from width to depth yields highly competitive task-specific mode
ls. Extensive experiments in two stages demonstrate the effectiveness of the MoP
E metric and MoPE-CLIP outperforms previous state-of-the-art VLP compression met
hods.
****
Learning Vision from Models Rivals Learning Vision from Data
Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, Phillip Iso
la; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2024, pp. 15887-15898
We introduce SynCLR a novel approach for learning visual representations exclusi
vely from synthetic images without any real data. We synthesize a large dataset
of image captions using LLMs then use an off-the-shelf text-to-image model to ge
nerate multiple images corresponding to each synthetic caption. We perform visua
l representation learning on these synthetic images via contrastive learning tre
ating images sharing the same caption as positive pairs. The resulting represent
ations demonstrate remarkable transferability competing favorably with other gen
eral-purpose visual representation learners such as CLIP and DINO v2 in image cl
assification tasks. Furthermore in dense prediction tasks such as semantic segme
ntation SynCLR outperforms previous self-supervised methods by a significant mar
gin e.g. improving over MAE and iBOT by 5.0 and 3.1 mIoU on ADE20k for ViT-B/16.
****
Leveraging Frame Affinity for sRGB-to-RAW Video De-rendering
Chen Zhang, Wencheng Han, Yang Zhou, Jianbing Shen, Cheng-zhong Xu, Wentao Liu;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2024, pp. 25659-25668
Unprocessed RAW video has shown distinct advantages over sRGB video in video edi
ting and computer vision tasks. However capturing RAW video is challenging due t
o limitations in bandwidth and storage. Various methods have been proposed to ad
dress similar issues in single image RAW capture through de-rendering. These met
hods utilize both the metadata and the sRGB image to perform sRGB-to-RAW de-rend
ering and recover high-quality single-frame RAW data. However metadata-based met
hods always require additional computation for online metadata generation imposi
ng severe burden on mobile camera device for high frame rate RAW video capture.
To address this issue we propose a framework that utilizes frame affinity to ach
ieve high-quality sRGB-to-RAW video reconstruction. Our approach consists of two
 main steps. The first step temporal affinity prior extraction uses motion infor
mation between adjacent frames to obtain a reference RAW image. The second step
spatial feature fusion and mapping learns a pixel-level mapping function using s
cene-specific and position-specific features provided by the previous frame. Our
 method can be easily applied to current mobile camera equipment without complic
ated adaptations or added burden. To demonstrate the effectiveness of our approa
ch we introduce the first RAW Video De-rendering Benchmark. In this benchmark ou
r method outperforms state-of-the-art RAW image reconstruction methods even with
out image-level metadata.
****
Adapting Short-Term Transformers for Action Detection in Untrimmed Videos
Min Yang, Huan Gao, Ping Guo, Limin Wang; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18570-18579
Vision Transformer (ViT) has shown high potential in video recognition owing to
its flexible design adaptable self-attention mechanisms and the efficacy of mask
ed pre-training. Yet it remains unclear how to adapt these pre-trained short-ter
m ViTs for temporal action detection (TAD) in untrimmed videos. The existing wor
ks treat them as off-the-shelf feature extractors for each short-trimmed snippet
 without capturing the fine-grained relation among different snippets in a broad
er temporal context. To mitigate this issue this paper focuses on designing a ne
w mechanism for adapting these pre-trained ViT models as a unified long-form vid
eo transformer to fully unleash its modeling power in capturing inter-snippet re
lation while still keeping low computation overhead and memory consumption for e
fficient TAD. To this end we design effective cross-snippet propagation modules

to gradually exchange short-term video information among different snippets from two levels. For inner-backbone information propagation we introduce a cross-snippet propagation strategy to enable multi-snippet temporal feature interaction inside the backbone. For post-backbone information propagation we propose temporal transformer layers for further clip-level modeling. With the plain ViT-B pre-trained with VideoMAE our end-to-end temporal action detector (ViT-TAD) yields a very competitive performance to previous temporal action detectors riching up to 69.5 average mAP on THUMOS14 37.40 average mAP on ActivityNet-1.3 and 17.20 average mAP on FineAction.
****
The Mirrored Influence Hypothesis: Efficient Data Influence Estimation by Harnessing Forward Passes
Myeongseob Ko, Feiyang Kang, Weiyan Shi, Ming Jin, Zhou Yu, Ruoxi Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26286-26295
Large-scale black-box models have become ubiquitous across numerous applications. Understanding the influence of individual training data sources on predictions made by these models is crucial for improving their trustworthiness. Current influence estimation techniques involve computing gradients for every training point or repeated training on different subsets. These approaches face obvious computational challenges when scaled up to large datasets and models. In this paper we introduce and explore the Mirrored Influence Hypothesis highlighting a reciprocal nature of influence between training and test data. Specifically it suggests that evaluating the influence of training data on test predictions can be reformulated as an equivalent yet inverse problem: assessing how the predictions for training samples would be altered if the model were trained on specific test samples. Through both empirical and theoretical validations we demonstrate the wide applicability of our hypothesis. Inspired by this we introduce a new method for estimating the influence of training data which requires calculating gradients for specific test samples paired with a forward pass for each training point. This approach can capitalize on the common asymmetry in scenarios where the number of test samples under concurrent examination is much smaller than the scale of the training dataset thus gaining a significant improvement in efficiency compared to existing approaches. We demonstrate the applicability of our method across a range of scenarios including data attribution in diffusion models data leakage detection analysis of memorization mislabeled data detection and tracing behavior in language models.
****
SOAC: Spatio-Temporal Overlap-Aware Multi-Sensor Calibration using Neural Radiance Fields
Quentin Herau, Nathan Piasco, Moussab Bennehar, Luis Roldao, Dzmitry Tsishkou, Cyrille Migniot, Pascal Vasseur, Cédric Demonceaux; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15131-15140
In rapidly-evolving domains such as autonomous driving the use of multiple sensors with different modalities is crucial to ensure high operational precision and stability. To correctly exploit the provided information by each sensor in a single common frame it is essential for these sensors to be accurately calibrated. In this paper we leverage the ability of Neural Radiance Fields (NeRF) to represent different sensors modalities in a common volumetric representation to achieve robust and accurate spatio-temporal sensor calibration. By designing a partitioning approach based on the visible part of the scene for each sensor we formulate the calibration problem using only the overlapping areas. This strategy results in a more robust and accurate calibration that is less prone to failure. We demonstrate that our approach works on outdoor urban scenes by validating it on multiple established driving datasets. Results show that our method is able to get better accuracy and robustness compared to existing methods.
****
G^3-LQ: Marrying Hyperbolic Alignment with Explicit Semantic-Geometric Modeling for 3D Visual Grounding

Yuan Wang, Yali Li, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13917-13926
Grounding referred objects in 3D scenes is a burgeoning vision-language task pivotal for propelling Embodied AI as it endeavors to connect the 3D physical world with free-form descriptions. Compared to the 2D counterparts challenges posed by the variability of 3D visual grounding remain relatively unsolved in existing studies: 1) the underlying geometric and complex spatial relationships in 3D scene. 2) the inherent complexity of 3D grounded language. 3) the inconsistencies between text and geometric features. To tackle these issues we propose G^3-LQ a DEtection TRansformer-based model tailored for 3D visual grounding task. G^3-LQ explicitly models Geometric-aware visual representations and Generates fine-Grained Language-guided object Queries in an overarching framework which comprises two dedicated modules. Specifically the Position Adaptive Geometric Exploring (PAGE) unearths underlying information of 3D objects in the geometric details and spatial relationships perspectives. The Fine-grained Language-guided Query Selection (Flan-QS) delves into syntactic structure of texts and generates object queries that exhibit higher relevance towards fine-grained text features. Finally a pioneering Poincare Semantic Alignment (PSA) loss establishes semantic-geometry consistencies by modeling non-linear vision-text feature mappings and aligning them on a hyperbolic prototype--Poincare ball. Extensive experiments verify the superiority of our G^3-LQ method trumping the state-of-the-arts by a considerable margin.
****
Garment Recovery with Shape and Deformation Priors
Ren Li, Corentin Dumery, Benoît Guillard, Pascal Fua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1586-1595
While modeling people wearing tight-fitting clothing has made great strides in recent years loose-fitting clothing remains a challenge. We propose a method that delivers realistic garment models from real-world images regardless of garment shape or deformation. To this end we introduce a fitting approach that utilizes shape and deformation priors learned from synthetic data to accurately capture garment shapes and deformations including large ones. Not only does our approach recover the garment geometry accurately it also yields models that can be directly used by downstream applications such as animation and simulation.
****
Psychometry: An Omnifit Model for Image Reconstruction from Human Brain Activity
Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 233-243
Reconstructing the viewed images from human brain activity bridges human and computer vision through the Brain-Computer Interface. The inherent variability in brain function between individuals leads existing literature to focus on acquiring separate models for each individual using their respective brain signal data ignoring commonalities between these data. In this article we devise Psychometry an omnifit model for reconstructing images from functional Magnetic Resonance Imaging (fMRI) obtained from different subjects. Psychometry incorporates an omni mixture-of-experts (Omni MoE) module where all the experts work together to capture the inter-subject commonalities while each expert associated with subject-specific parameters copes with the individual differences. Moreover Psychometry is equipped with a retrieval-enhanced inference strategy termed Ecphory which aims to enhance the learned fMRI representation via retrieving from prestored subject-specific memories. These designs collectively render Psychometry omnifit and efficient enabling it to capture both inter-subject commonality and individual specificity across subjects. As a result the enhanced fMRI representations serve as conditional signals to guide a generation model to reconstruct high-quality and realistic images establishing Psychometry as state-of-the-art in terms of both high-level and low-level metrics.
****
Exploring Regional Clues in CLIP for Zero-Shot Semantic Segmentation

Yi Zhang, Meng-Hao Guo, Miao Wang, Shi-Min Hu; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3270-3280
CLIP has demonstrated marked progress in visual recognition due to its powerful pre-training on large-scale image-text pairs. However it still remains a critica l challenge: how to transfer image-level knowledge into pixel-level understandin g tasks such as semantic segmentation. In this paper to solve the mentioned chal lenge we analyze the gap between the capability of the CLIP model and the requir ement of the zero-shot semantic segmentation task. Based on our analysis and obs ervations we propose a novel method for zero-shot semantic segmentation dubbed C LIP-RC (CLIP with Regional Clues) bringing two main insights. On the one hand a region-level bridge is necessary to provide fine-grained semantics. On the other hand overfitting should be mitigated during the training stage. Benefiting from the above discoveries CLIP-RC achieves state-of-the-art performance on various zero-shot semantic segmentation benchmarks including PASCAL VOC PASCAL Context a nd COCO-Stuff 164K. Code will be available at https://github.com/Jittor/JSeg.
****
Move as You Say Interact as You Can: Language-guided Human Motion Generation wit h Scene Affordance
Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, Siyuan Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 433-444
Despite significant advancements in text-to-motion synthesis generating language -guided human motion within 3D environments poses substantial challenges. These challenges stem primarily from (i) the absence of powerful generative models cap able of jointly modeling natural language 3D scenes and human motion and (ii) th e generative models' intensive data requirements contrasted with the scarcity of comprehensive high-quality language-scene-motion datasets. To tackle these issu es we introduce a novel two-stage framework that employs scene affordance as an intermediate representation effectively linking 3D scene grounding and condition al motion generation. Our framework comprises an Affordance Diffusion Model (ADM) for predicting explicit affordance map and an Affordance-to-Motion Diffusion M odel (AMDM) for generating plausible human motions. By leveraging scene affordan ce maps our method overcomes the difficulty in generating human motion under mul timodal condition signals especially when training with limited data lacking ext ensive language-scene-motion pairs. Our extensive experiments demonstrate that o ur approach consistently outperforms all baselines on established benchmarks inc luding HumanML3D and HUMANISE. Additionally we validate our model's exceptional generalization capabilities on a specially curated evaluation set featuring prev iously unseen descriptions and scenes.
****
Choose What You Need: Disentangled Representation Learning for Scene Text Recogn ition Removal and Editing
Boqiang Zhang, Hongtao Xie, Zuan Gao, Yuxin Wang; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28358-2836 8
Scene text images contain not only style information (font background) but also content information (character texture). Different scene text tasks need differe nt information but previous representation learning methods use tightly coupled features for all tasks resulting in sub-optimal performance. We propose a Disent angled Representation Learning framework (DARLING) aimed at disentangling these two types of features for improved adaptability in better addressing various dow nstream tasks (choose what you really need). Specifically we synthesize a datase t of image pairs with identical style but different content. Based on the datase t we decouple the two types of features by the supervision design. Clearly we di rectly split the visual representation into style and content features the conte nt features are supervised by a text recognition loss while an alignment loss al igns the style features in the image pairs. Then style features are employed in reconstructing the counterpart image via an image decoder with a prompt that ind icates the counterpart's content. Such an operation effectively decouples the fe atures based on their distinctive properties. To the best of our knowledge this

is the first time in the field of scene text that disentangles the inherent prop
erties of the text images. Our method achieves state-of-the-art performance in S
cene Text Recognition Removal and Editing.
****
Generalizable Face Landmarking Guided by Conditional Face Warping
Jiayi Liang, Haotian Liu, Hongteng Xu, Dixin Luo; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2425-2435
As a significant step for human face modeling editing and generation face landma
rking aims at extracting facial keypoints from images. A generalizable face land
marker is required in practice because real-world facial images e.g. the avatars
 in animations and games are often stylized in various ways. However achieving g
eneralizable face landmarking is challenging due to the diversity of facial styl
es and the scarcity of labeled stylized faces. In this study we propose a simple
 but effective paradigm to learn a generalizable face landmarker based on labele
d real human faces and unlabeled stylized faces. Our method learns the face land
marker as the key module of a conditional face warper. Given a pair of real and
stylized facial images the conditional face warper predicts a warping field from
 the real face to the stylized one in which the face landmarker predicts the end
ing points of the warping field and provides us with high-quality pseudo landmar
ks for the corresponding stylized facial images. Applying an alternating optimiz
ation strategy we learn the face landmarker to minimize i) the discrepancy betwe
en the stylized faces and the warped real ones and ii) the prediction errors of
both real and pseudo landmarks. Experiments on various datasets show that our me
thod outperforms existing state-of-the-art domain adaptation methods in face lan
dmarking tasks leading to a face landmarker with better generalizability. Code i
s available at https://plustwo0.github.io/project-face-landmarker.
****
Sat2Scene: 3D Urban Scene Generation from Satellite Images with Diffusion
Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Marc Pollefeys, Martin R. Oswald; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 7141-7150
Directly generating scenes from satellite imagery offers exciting possibilities
for integration into applications like games and map services. However challenge
s arise from significant view changes and scene scale. Previous efforts mainly f
ocused on image or video generation lacking exploration into the adaptability of
 scene generation for arbitrary views. Existing 3D generation works either opera
te at the object level or are difficult to utilize the geometry obtained from sa
tellite imagery. To overcome these limitations we propose a novel architecture f
or direct 3D scene generation by introducing diffusion models into 3D sparse rep
resentations and combining them with neural rendering techniques. Specifically o
ur approach generates texture colors at the point level for a given geometry usi
ng a 3D diffusion model first which is then transformed into a scene representat
ion in a feed-forward manner. The representation can be utilized to render arbit
rary views which would excel in both single-frame quality and inter-frame consis
tency. Experiments in two city-scale datasets show that our model demonstrates p
roficiency in generating photo-realistic street-view image sequences and cross-v
iew urban scenes from satellite imagery.
****
Control4D: Efficient 4D Portrait Editing with Text
Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang,
 Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2024, pp. 4556-4567
We introduce Control4D an innovative framework for editing dynamic 4D portraits
using text instructions. Our method addresses the prevalent challenges in 4D edi
ting notably the inefficiencies of existing 4D representations and the inconsist
ent editing effect caused by diffusion-based editors. We first propose GaussianP
lanes a novel 4D representation that makes Gaussian Splatting more structured by
 applying plane-based decomposition in 3D space and time. This enhances both eff
iciency and robustness in 4D editing. Furthermore we propose to leverage a 4D ge
nerator to learn a more continuous generation space from inconsistent edited ima

ges produced by the diffusion-based editor which effectively improves the consis tency and quality of 4D editing. Comprehensive evaluation demonstrates the super iority of Control4D including significantly reduced training time high-quality r endering and spatial-temporal consistency in 4D portrait editing.
****
Symphonize 3D Semantic Scene Completion with Contextual Instance Queries
Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, X inggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2024, pp. 20258-20267
3D Semantic Scene Completion (SSC) has emerged as a nascent and pivotal undertak ing in autonomous driving aiming to predict the voxel occupancy within volumetri c scenes. However prevailing methodologies primarily focus on voxel-wise feature aggregation while neglecting instance semantics and scene context. In this pape r we present a novel paradigm termed Symphonies (Scene-from-Insts) that delves i nto the integration of instance queries to orchestrate 2D-to-3D reconstruction a nd 3D scene modeling. Leveraging our proposed Serial Instance-Propagated Attenti ons Symphonies dynamically encodes instance-centric semantics facilitating intri cate interactions between the image and volumetric domains. Simultaneously Symph onies fosters holistic scene comprehension by capturing context through the effi cient fusion of instance queries alleviating geometric ambiguities such as occlu sion and perspective errors through contextual scene reasoning. Experimental res ults demonstrate that Symphonies achieves state-of-the-art performance on the ch allenging SemanticKITTI and SSCBench-KITTI-360 benchmarks yielding remarkable mI oU scores of 15.04 and 18.58 respectively. These results showcase the promising advancements of our paradigm. The code for our method is available at https://gi thub.com/hustvl/Symphonies.
****
Loopy-SLAM: Dense Neural SLAM with Loop Closures
Lorenzo Liso, Erik Sandström, Vladimir Yugay, Luc Van Gool, Martin R. Oswald; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20363-20373
Neural RGBD SLAM techniques have shown promise in dense Simultaneous Localizatio n And Mapping (SLAM) yet face challenges such as error accumulation during camer a tracking resulting in distorted maps. In response we introduce Loopy-SLAM that globally optimizes poses and the dense 3D model. We use frame-to-model tracking using a data-driven point-based submap generation method and trigger loop closu res online by performing global place recognition. Robust pose graph optimizatio n is used to rigidly align the local submaps. As our representation is point bas ed map corrections can be performed efficiently without the need to store the en tire history of input frames used for mapping as typically required by methods e mploying a grid based mapping structure. Evaluation on the synthetic Replica and real-world TUM-RGBD and ScanNet datasets demonstrate competitive or superior pe rformance in tracking mapping and rendering accuracy when compared to existing d ense neural RGBD SLAM methods. Project page: notchla.github.io/Loopy-SLAM.
****
CLIPtone: Unsupervised Learning for Text-based Image Tone Adjustment
Hyeongmin Lee, Kyoungkook Kang, Jungseul Ok, Sunghyun Cho; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2 942-2951
Recent image tone adjustment (or enhancement) approaches have predominantly adop ted supervised learning for learning human-centric perceptual assessment. Howeve r these approaches are constrained by intrinsic challenges of supervised learnin g. Primarily the requirement for expertly-curated or retouched images escalates the data acquisition expenses. Moreover their coverage of target styles is confi ned to stylistic variants inferred from the training data. To surmount the above challenges we propose an unsupervised learning-based approach for text-based im age tone adjustment CLIPtone that extends an existing image enhancement method t o accommodate natural language descriptions. Specifically we design a hyper-netw ork to adaptively modulate the pretrained parameters of a backbone model based o n a text description. To assess whether an adjusted image aligns with its text d

escription without a ground-truth image we utilize CLIP which is trained on a va
st set of language-image pairs and thus encompasses the knowledge of human perce
ption. The major advantages of our approach are threefold: (i) minimal data coll
ection expenses (ii) support for a range of adjustments and (iii) the ability to
 handle novel text descriptions unseen in training. The efficacy of the proposed
 method is demonstrated through comprehensive experiments including a user study
.
****
ToonerGAN: Reinforcing GANs for Obfuscating Automated Facial Indexing
Kartik Thakral, Shashikant Prasad, Stuti Aswani, Mayank Vatsa, Richa Singh; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 10875-10884
The rapid evolution of automatic facial indexing tech- nologies increases the ri
sk of compromising personal and sensitive information. To address the issue we p
ropose cre- ating cartoon avatars or 'toon avatars' designed to effec- tively ob
scure identity features. The primary objective is to deceive current AI systems
preventing them from accu- rately identifying individuals while making minimal m
odi- fications to their facial features. Moreover we aim to en- sure that a huma
n observer can still recognize the person depicted in these altered avatar image
s. To achieve this we introduce 'ToonerGAN' a novel approach that utilizes Gener
ative Adversarial Networks (GANs) to craft person- alized cartoon avatars. The T
oonerGAN framework con- sists of a style module and a de-identification module t
hat work together to produce high-resolution realistic cartoon images. For the e
fficient training of our network we have developed an extensive dataset named 'T
oonSet' compris- ing approximately 23000 facial images and their cartoon renditi
ons. Through comprehensive experiments and bench- marking against existing datas
ets including CelebA-HQ our method demonstrates superior performance in obfus- c
ating identity while preserving the utility of data. Addi- tionally a user-centr
ic study to explore the effectiveness of ToonerGAN has yielded some compelling o
bservations.
****
Content-Adaptive Non-Local Convolution for Remote Sensing Pansharpening
Yule Duan, Xiao Wu, Haoyu Deng, Liang-Jian Deng; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27738-27747
Currently machine learning-based methods for remote sensing pansharpening have p
rogressed rapidly. However existing pansharpening methods often do not fully exp
loit differentiating regional information in non-local spaces thereby limiting t
he effectiveness of the methods and resulting in redundant learning parameters.
In this paper we introduce a so-called content-adaptive non-local convolution (C
ANConv) a novel method tailored for remote sensing image pansharpening. Specific
ally CANConv employs adaptive convolution ensuring spatial adaptability and inco
rporates non-local self-similarity through the similarity relationship partition
 (SRP) and the partition-wise adaptive convolution (PWAC) sub-modules. Furthermo
re we also propose a corresponding network architecture called CANNet which main
ly utilizes the multi-scale self-similarity. Extensive experiments demonstrate t
he superior performance of CANConv compared with recent promising fusion methods
. Besides we substantiate the method's effectiveness through visualization ablat
ion experiments and comparison with existing methods on multiple test sets. The
source code is publicly available at https://github.com/duanyll/CANConv.
****
Codebook Transfer with Part-of-Speech for Vector-Quantized Image Modeling
Baoquan Zhang, Huaibin Wang, Chuyao Luo, Xutao Li, Guotao Liang, Yunming Ye, Xia
ochen Qi, Yao He; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2024, pp. 7757-7766
Vector-Quantized Image Modeling (VQIM) is a fundamental research problem in imag
e synthesis which aims to represent an image with a discrete token sequence. Exi
sting studies effectively address this problem by learning a discrete codebook f
rom scratch and in a code-independent manner to quantize continuous representati
ons into discrete tokens. However learning a codebook from scratch and in a code
-independent manner is highly challenging which may be a key reason causing code

book collapse i.e. some code vectors can rarely be optimized without regard to the relationship between codes and good codebook priors such that die off finally. In this paper inspired by pretrained language models we find that these language models have actually pretrained a superior codebook via a large number of text corpus but such information is rarely exploited in VQIM. To this end we propose a novel codebook transfer framework with part-of-speech called VQCT which aims to transfer a well-trained codebook from pretrained language models to VQIM for robust codebook learning. Specifically we first introduce a pretrained codebook from language models and part-of-speech knowledge as priors. Then we construct a vision-related codebook with these priors for achieving codebook transfer. Finally a novel codebook transfer network is designed to exploit abundant semantic relationships between codes contained in pretrained codebooks for robust VQIM codebook learning. Experimental results on four datasets show that our VQCT method achieves superior VQIM performance over previous state-of-the-art methods.
****
Learning Inclusion Matching for Animation Paint Bucket Colorization
Yuekun Dai, Shangchen Zhou, Qinyue Li, Chongyi Li, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25544-25553
Colorizing line art is a pivotal task in the production of hand-drawn cel animation. This typically involves digital painters using a paint bucket tool to manually color each segment enclosed by lines based on RGB values predetermined by a color designer. This frame-by-frame process is both arduous and time-intensive. Current automated methods mainly focus on segment matching. This technique migrates colors from a reference to the target frame by aligning features within line-enclosed segments across frames. However issues like occlusion and wrinkles in animations often disrupt these direct correspondences leading to mismatches. In this work we introduce a new learning-based inclusion matching pipeline which directs the network to comprehend the inclusion relationships between segments rather than relying solely on direct visual correspondences. Our method features a two-stage pipeline that integrates a coarse color warping module with an inclusion matching module enabling more nuanced and accurate colorization. To facilitate the training of our network we also develope a unique dataset referred to as PaintBucket-Character. This dataset includes rendered line arts alongside their colorized counterparts featuring various 3D characters. Extensive experiments demonstrate the effectiveness and superiority of our method over existing techniques.
****
Editable Scene Simulation for Autonomous Driving via Collaborative LLM-Agents
Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, Yanfeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15077-15087
Scene simulation in autonomous driving has gained significant attention because of its huge potential for generating customized data. However existing editable scene simulation approaches face limitations in terms of user interaction efficiency multi-camera photo-realistic rendering and external digital assets integration. To address these challenges this paper introduces ChatSim the first system that enables editable photo-realistic 3D driving scene simulations via natural language commands with external digital assets. To enable editing with high command flexibility ChatSim leverages a large language model (LLM) agent collaboration framework. To generate photo-realistic outcomes ChatSim employs a novel multi-camera neural radiance field method. Furthermore to unleash the potential of extensive high-quality digital assets ChatSim employs a novel multi-camera lighting estimation method to achieve scene-consistent assets' rendering. Our experiments on Waymo Open Dataset demonstrate that ChatSim can handle complex language commands and generate corresponding photo-realistic scene videos. Code can be accessed at: https://github.com/yifanlu0227/ChatSim.
****
SAM-6D: Segment Anything Model Meets Zero-Shot 6D Object Pose Estimation
Jiehong Lin, Lihua Liu, Dekun Lu, Kui Jia; Proceedings of the IEEE/CVF Conferenc

e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27906-27916
Zero-shot 6D object pose estimation involves the detection of novel objects with their 6D poses in cluttered scenes presenting significant challenges for model generalizability. Fortunately the recent Segment Anything Model (SAM) has showcased remarkable zero-shot transfer performance which provides a promising solution to tackle this task. Motivated by this we introduce SAM-6D a novel framework designed to realize the task through two steps including instance segmentation and pose estimation. Given the target objects SAM-6D employs two dedicated sub-networks namely Instance Segmentation Model (ISM) and Pose Estimation Model (PEM) to perform these steps on cluttered RGB-D images. ISM takes SAM as an advanced starting point to generate all possible object proposals and selectively preserves valid ones through meticulously crafted object matching scores in terms of semantics appearance and geometry. By treating pose estimation as a partial-to-partial point matching problem PEM performs a two-stage point matching process featuring a novel design of background tokens to construct dense 3D-3D correspondence ultimately yielding the pose estimates. Without bells and whistles SAM-6D outperforms the existing methods on the seven core datasets of the BOP Benchmark for both instance segmentation and pose estimation of novel objects.
****
InceptionNeXt: When Inception Meets ConvNeXt
Weihao Yu, Pan Zhou, Shuicheng Yan, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5672-5683
Inspired by the long-range modeling ability of ViTs large-kernel convolutions are widely studied and adopted recently to enlarge the receptive field and improve model performance like the remarkable work ConvNeXt which employs 7x7 depthwise convolution. Although such depthwise operator only consumes a few FLOPs it largely harms the model efficiency on powerful computing devices due to the high memory access costs. For example ConvNeXt-T has similar FLOPs with ResNet-50 but only achieves 60% throughputs when trained on A100 GPUs with full precision. Although reducing the kernel size of ConvNeXt can improve speed it results in significant performance degradation which poses a challenging problem: How to speed up large-kernel-based CNN models while preserving their performance. To tackle this issue inspired by Inceptions we propose to decompose large-kernel depthwise convolution into four parallel branches along channel dimension i.e. small square kernel two orthogonal band kernels and an identity mapping. With this new Inception depthwise convolution we build a series of networks namely IncepitonNeXt which not only enjoy high throughputs but also maintain competitive performance. For instance InceptionNeXt-T achieves 1.6x higher training throughputs than ConvNeX-T as well as attains 0.2% top-1 accuracy improvement on ImageNet-1K. We anticipate InceptionNeXt can serve as an economical baseline for future architecture design to reduce carbon footprint.
****
SnAG: Scalable and Accurate Video Grounding
Fangzhou Mu, Sicheng Mo, Yin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18930-18940
Temporal grounding of text descriptions in videos is a central problem in vision-language learning and video understanding. Existing methods often prioritize accuracy over scalability --- they have been optimized for grounding only a few text queries within short videos and fail to scale up to long videos with hundreds of queries. In this paper we study the effect of cross-modal fusion on the scalability of video grounding models. Our analysis establishes late fusion as a more cost-effective fusion scheme for long-form videos with many text queries. Moreover it leads us to a novel video-centric sampling scheme for efficient training. Based on these findings we present SnAG a simple baseline for scalable and accurate video grounding. Without bells and whistles SnAG is 43% more accurate and 1.5x faster than CONE a state of the art for long-form video grounding on the challenging MAD dataset while achieving highly competitive results on short videos.
****
SPOT: Self-Training with Patch-Order Permutation for Object-Centric Learning wit

h Autoregressive Transformers
Ioannis Kakogeorgiou, Spyros Gidaris, Konstantinos Karantzalos, Nikos Komodakis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22776-22786
Unsupervised object-centric learning aims to decompose scenes into interpretable object entities termed slots. Slot-based auto-encoders stand out as a prominent method for this task. Within them crucial aspects include guiding the encoder to generate object-specific slots and ensuring the decoder utilizes them during reconstruction. This work introduces two novel techniques (i) an attention-based self-training approach which distills superior slot-based attention masks from the decoder to the encoder enhancing object segmentation and (ii) an innovative patch-order permutation strategy for autoregressive transformers that strengthens the role of slot vectors in reconstruction. The effectiveness of these strategies is showcased experimentally. The combined approach significantly surpasses prior slot-based autoencoder methods in unsupervised object segmentation especially with complex real-world images. We provide the implementation code at https://github.com/gkakogeorgiou/spot .
****
LiveHPS: LiDAR-based Scene-level Human Pose and Shape Estimation in Free Environment
Yiming Ren, Xiao Han, Chengfeng Zhao, Jingya Wang, Lan Xu, Jingyi Yu, Yuexin Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1281-1291
For human-centric large-scale scenes fine-grained modeling for 3D human global pose and shape is significant for scene understanding and can benefit many real-world applications. In this paper we present LiveHPS a novel single-LiDAR-based approach for scene-level human pose and shape estimation without any limitation of light conditions and wearable devices. In particular we design a distillation mechanism to mitigate the distribution-varying effect of LiDAR point clouds and exploit the temporal-spatial geometric and dynamic information existing in consecutive frames to solve the occlusion and noise disturbance. LiveHPS with its efficient configuration and high-quality output is well-suited for real-world applications. Moreover we propose a huge human motion dataset named FreeMotion which is collected in various scenarios with diverse human poses shapes and translations. It consists of multi-modal and multi-view acquisition data from calibrated and synchronized LiDARs cameras and IMUs. Extensive experiments on our new dataset and other public datasets demonstrate the SOTA performance and robustness of our approach. We will release our code and dataset soon.
****
Segment Every Out-of-Distribution Object
Wenjie Zhao, Jia Li, Xin Dong, Yu Xiang, Yunhui Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3910-3920
Semantic segmentation models while effective for in-distribution categories face challenges in real-world deployment due to encountering out-of-distribution (OoD) objects. Detecting these OoD objects is crucial for safety-critical applications. Existing methods rely on anomaly scores but choosing a suitable threshold for generating masks presents difficulties and can lead to fragmentation and inaccuracy. This paper introduces a method to convert anomaly Score To segmentation Mask called S2M a simple and effective framework for OoD detection in semantic segmentation. Unlike assigning anomaly scores to pixels S2M directly segments the entire OoD object. By transforming anomaly scores into prompts for a promptable segmentation model S2M eliminates the need for threshold selection. Extensive experiments demonstrate that S2M outperforms the state-of-the-art by approximately 20% in IoU and 40% in mean F1 score on average across various benchmarks including Fishyscapes Segment-Me-If-You-Can and RoadAnomaly datasets.
****
Building Vision-Language Models on Solid Foundations with Masked Distillation
Sepehr Sameni, Kushal Kafle, Hao Tan, Simon Jenni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14216-142

Recent advancements in Vision-Language Models (VLMs) have marked a significant leap in bridging the gap between computer vision and natural language processing. However traditional VLMs trained through contrastive learning on limited and noisy image-text pairs often lack the spatial and linguistic understanding to generalize well to dense vision tasks or less common languages. Our approach Solid Foundation CLIP (SF-CLIP) circumvents this issue by implicitly building on the solid visual and language understanding of foundational models trained on vast amounts of unimodal data. SF-CLIP integrates contrastive image-text pretraining with a masked knowledge distillation from large foundational text and vision models. This methodology guides our VLM in developing robust text and image representations. As a result SF-CLIP shows exceptional zero-shot classification accuracy and enhanced image and text retrieval capabilities setting a new state of the art for ViT-B/16 trained on YFCC15M and CC12M. Moreover the dense per-patch supervision enhances our zero-shot and linear probe performance in semantic segmentation tasks. A remarkable aspect of our model is its multilingual proficiency evidenced by strong retrieval results in multiple languages despite being trained predominantly on English data. We achieve all of these improvements without sacrificing the training efficiency through our selective application of masked distillation and the inheritance of teacher word embeddings.
****
Wavelet-based Fourier Information Interaction with Frequency Diffusion Adjustment for Underwater Image Restoration
Chen Zhao, Weiling Cai, Chenyu Dong, Chengwei Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8281-8291
Underwater images are subject to intricate and diverse degradation inevitably affecting the effectiveness of underwater visual tasks. However most approaches primarily operate in the raw pixel space of images which limits the exploration of the frequency characteristics of underwater images leading to an inadequate utilization of deep models' representational capabilities in producing high-quality images. In this paper we introduce a novel Underwater Image Enhancement (UIE) framework named WF-Diff designed to fully leverage the characteristics of frequency domain information and diffusion models. WF-Diff consists of two detachable networks: Wavelet-based Fourier information interaction network (WFI2-net) and Frequency Residual Diffusion Adjustment Module (FRDAM). With our full exploration of the frequency domain information WFI2-net aims to achieve preliminary enhancement of frequency information in the wavelet space. Our proposed FRDAM can further refine the high- and low-frequency information of the initial enhanced images which can be viewed as a plug-and-play universal module to adjust the detail of the underwater images. With the above techniques our algorithm can show SOTA performance on real-world underwater image datasets and achieves competitive performance in visual quality.
****
CroSel: Cross Selection of Confident Pseudo Labels for Partial-Label Learning
Shiyu Tian, Hongxin Wei, Yiqun Wang, Lei Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19479-19488
Partial-label learning (PLL) is an important weakly supervised learning problem which allows each training example to have a candidate label set instead of a single ground-truth label. Identification-based methods have been widely explored to tackle label ambiguity issues in PLL which regard the true label as a latent variable to be identified. However identifying the true labels accurately and completely remains challenging causing noise in pseudo labels during model training. In this paper we propose a new method called CroSel which leverages historical predictions from the model to identify true labels for most training examples. First we introduce a cross selection strategy which enables two deep models to select true labels of partially labeled data for each other. Besides we propose a novel consistency regularization term called co-mix to avoid sample waste and tiny noise caused by false selection. In this way CroSel can pick out the true labels of most examples with high precision. Extensive experiments demonstrate the superiority of CroSel which consistently outperforms previous state-of-the-art

methods on benchmark datasets. Additionally our method achieves over 90% accura cy and quantity for selecting true labels on CIFAR-type datasets under various s ettings.
****

PoNQ: a Neural QEM-based Mesh Representation
Nissim Maruani, Maks Ovsjanikov, Pierre Alliez, Mathieu Desbrun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3647-3657
Although polygon meshes have been a standard representation in geometry processi ng their irregular and combinatorial nature hinders their suitability for learni ng-based applications. In this work we introduce a novel learnable mesh represen tation through a set of local 3D sample Points and their associated Normals and Quadric error metrics (QEM) w.r.t. the underlying shape which we denote PoNQ. A global mesh is directly derived from PoNQ by efficiently leveraging the knowledg e of the local quadric errors. Besides marking the first use of QEM within a neu ral shape representation our contribution guarantees both topological and geomet rical properties by ensuring that a PoNQ mesh does not self-intersect and is alw ays the boundary of a volume. Notably our representation does not rely on a regu lar grid is supervised directly by the target surface alone and also handles ope n surfaces with boundaries and/or sharp features. We demonstrate the efficacy of PoNQ through a learning-based mesh prediction from SDF grids and show that our method surpasses recent state-of-the-art techniques in terms of both surface and edge-based metrics.
****

ModaVerse: Efficiently Transforming Modalities with LLMs
Xinyu Wang, Bohan Zhuang, Qi Wu; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2024, pp. 26606-26616
Humans possess the capability to comprehend diverse modalities and seamlessly tr ansfer information between them. In this work we introduce ModaVerse a Multi-mod al Large Language Model (MLLM) capable of comprehending and transforming content across various modalities including images videos and audio. Predominant MLLM f rameworks have largely relied on aligning latent spaces of textual and non-textu al features. This alignment process which synchronizes a language model trained on textual data with encoders and decoders trained on multi-modal data often nec essitates extensive training of several projection layers in multiple stages. In spired by LLM-as-agent methodologies we propose a novel Input/Output (I/O) align ment mechanism that operates directly at the level of natural language. It align s the LLM's output with the input of generative models avoiding the complexities associated with latent feature alignments and simplifying the multiple training stages of existing MLLMs into a single efficient process. By conducting experim ents on several benchmarks we demonstrate that our approach attains comparable p erformance with the state of the art while achieving considerable efficiencies i n data usage.
****

TransLoc4D: Transformer-based 4D Radar Place Recognition
Guohao Peng, Heshan Li, Yangyang Zhao, Jun Zhang, Zhenyu Wu, Pengyu Zheng, Danwe i Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2024, pp. 17595-17605
Place Recognition is crucial for unmanned vehicles in terms of localization and mapping. Recent years have witnessed numerous explorations in the field where 2D cameras and 3D LiDARs are mostly employed. Despite their admirable performance they may encounter challenges in adverse weather such as rain and fog. Hopefully 4D millimeter-wave Radar emerges as a promising alternative as its longer wavel ength makes it virtually immune to interference from tiny particles of fog and r ain. Therefore in this work we propose a novel 4D Radar place recognition model TransLoc4D based on sparse convolution and Transformer structures. Specifically a Minkloc4D backbone is first proposed to leverage the geometric intensity and v elocity information from 4D Radar scans. While mainstream 3D LiDAR solutions mer ely capture geometric structures of point clouds Minkloc4D explores the intensit y and velocity properties of 4D Radar scans and demonstrates their effectiveness

. After feature extraction a Transformer layer is introduced to enhance local features where linear self-attention captures the long-range dependency of point cloud alleviating its sparsity and noise. To validate TransLoc4D we construct two datasets and set up benchmarks for 4D radar place recognition. Experiments show TransLoc4D is feasible and can robustly deal with dynamic and adverse environments.
****
Frequency-aware Event-based Video Deblurring for Real-World Motion Blur
Taewoo Kim, Hoonhee Cho, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24966-24976
Video deblurring aims to restore sharp frames from blurred video clips. Despite notable progress in video deblurring works it is still a challenging problem because of the loss of motion information during the duration of the exposure time. Since event cameras can capture clear motion information asynchronously with high temporal resolution several works exploit the event camera for deblurring as they can provide abundant motion information. However despite these approaches there were few cases of actively exploiting the long-range temporal dependency of videos. To tackle these deficiencies we present an event-based video deblurring framework by actively utilizing temporal information from videos. To be specific we first introduce a frequency-based cross-modal feature enhancement module. Second we propose event-guided video alignment modules by considering the valuable characteristics of the event and videos. In addition we designed a hybrid camera system to collect the first real-world event-based video deblurring dataset. For the first time we build a dataset containing synchronized high-resolution real-world blurred videos and corresponding sharp videos and event streams. Experimental results validate that our frameworks significantly outperform the state-of-the-art frame-based and event-based deblurring works in the various datasets. In addition we designed a hybrid camera system to collect the first real-world event-based video deblurring dataset. For the first time we build a dataset containing synchronized high-resolution real-world blurred videos and corresponding sharp videos and event streams. Experimental results validate that our frameworks significantly outperform the state-of-the-art frame-based and event-based deblurring works in the various datasets. The project pages are available at https://sites.google.com/view/fevd-cvpr2024.
****
Multiscale Vision Transformers Meet Bipartite Matching for Efficient Single-stage Action Localization
Ioanna Ntinou, Enrique Sanchez, Georgios Tzimiropoulos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18827-18836
Action Localization is a challenging problem that combines detection and recognition tasks which are often addressed separately. State-of-the-art methods rely on off-the-shelf bounding box detections pre-computed at high resolution and propose transformer models that focus on the classification task alone. Such two-stage solutions are prohibitive for real-time deployment. On the other hand single-stage methods target both tasks by devoting part of the network (generally the backbone) to sharing the majority of the workload compromising performance for speed. These methods build on adding a DETR head with learnable queries that after cross- and self-attention can be sent to corresponding MLPs for detecting a person's bounding box and action. However DETR-like architectures are challenging to train and can incur in big complexity. In this paper we observe that a straight bipartite matching loss can be applied to the output tokens of a vision transformer. This results in a backbone + MLP architecture that can do both tasks without the need of an extra encoder-decoder head and learnable queries. We show that a single MViTv2-S architecture trained with bipartite matching to perform both tasks surpasses the same MViTv2-S when trained with RoI align on pre-computed bounding boxes. With a careful design of token pooling and the proposed training pipeline our Bipartite-Matching Vision Transformer model BMViT achieves +3 mAP on AVA2.2. w.r.t. the two-stage MViTv2-S counterpart. Code is available at https://github.com/IoannaNti/BMViT

****
Boosting Order-Preserving and Transferability for Neural Architecture Search: a Joint Architecture Refined Search and Fine-tuning Approach
Beichen Zhang, Xiaoxing Wang, Xiaohan Qin, Junchi Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5662-5671
Supernet is a core component in many recent Neural Architecture Search (NAS) methods. It not only helps embody the search space but also provides a (relative) estimation of the final performance of candidate architectures. Thus it is critical that the top architectures ranked by a supernet should be consistent with those ranked by true performance which is known as the order-preserving ability. In this work we analyze the order-preserving ability on the whole search space (global) and a sub-space of top architectures (local) and empirically show that the local order-preserving for current two-stage NAS methods still need to be improved. To rectify this we propose a novel concept of Supernet Shifting a refined search strategy combining architecture searching with supernet fine-tuning. Specifically apart from evaluating the training loss is also accumulated in searching and the supernet is updated every iteration. Since superior architectures are sampled more frequently in evolutionary searching the supernet is encouraged to focus on top architectures thus improving local order-preserving. Besides a pre-trained supernet is often un-reusable for one-shot methods. We show that Supernet Shifting can fulfill transferring supernet to a new dataset. Specifically the last classifier layer will be unset and trained through evolutionary searching. Comprehensive experiments show that our method has better order-preserving ability and can find a dominating architecture. Moreover the pre-trained supernet can be easily transferred into a new dataset with no loss of performance.
****
Dr. Bokeh: DiffeRentiable Occlusion-aware Bokeh Rendering
Yichen Sheng, Zixun Yu, Lu Ling, Zhiwen Cao, Xuaner Zhang, Xin Lu, Ke Xian, Haiting Lin, Bedrich Benes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4515-4525
Bokeh is widely used in photography to draw attention to the subject while effectively isolating distractions in the background. Computational methods can simulate bokeh effects without relying on a physical camera lens but the inaccurate lens modeling in existing filtering-based methods leads to artifacts that need post-processing or learning-based methods to fix. We propose Dr.Bokeh a novel rendering method that addresses the issue by directly correcting the defect that violates the physics in the current filtering-based bokeh rendering equation. Dr.Bokeh first preprocesses the input RGBD to obtain a layered scene representation. Dr.Bokeh then takes the layered representation and user-defined lens parameters to render photo-realistic lens blur based on the novel occlusion-aware bokeh rendering method. Experiments show that the non-learning based renderer Dr.Bokeh outperforms state-of-the-art bokeh rendering algorithms in terms of photo-realism. In addition extensive quantitative and qualitative evaluations show the more accurate lens model further pushes the limit of a closely related field depth-from-defocus.
****
Unsegment Anything by Simulating Deformation
Jiahao Lu, Xingyi Yang, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24294-24304
Foundation segmentation models while powerful pose a significant risk: they enable users to effortlessly extract any objects from any digital content with a single click potentially leading to copyright infringement or malicious misuse. To mitigate this risk we introduce a new task "Anything Unsegmentable" to grant any image "the right to be unsegmented". The ambitious pursuit of the task is to achieve highly transferable adversarial attack against all prompt-based segmentation models regardless of model parameterizations and prompts. We highlight the non-transferable and heterogeneous nature of prompt-specific adversarial noises. Our approach focuses on disrupting image encoder features to achieve prompt-agnostic attacks. Intriguingly targeted feature attacks exhibit better transferabilit

y compared to untargeted ones suggesting the optimal update direction aligns with the image manifold. Based on the observations we design a novel attack named Unsegment Anything by Simulating Deformation (UAD). Our attack optimizes a differentiable deformation function to create a target deformed image which alters structural information while preserving achievable feature distance by adversarial example. Extensive experiments verify the effectiveness of our approach compromising a variety of promptable segmentation models with different architectures and prompt interfaces.
****
Transductive Zero-Shot and Few-Shot CLIP
Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, Ismail Ben Ayed; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28816-28826
Transductive inference has been widely investigated in few-shot image classification but completely overlooked in the recent fast growing literature on adapting vision-langage models like CLIP. This paper addresses the transductive zero-shot and few-shot CLIP classification challenge in which inference is performed jointly across a mini-batch of unlabeled query samples rather than treating each instance independently. This paper addresses the transductive zero-shot and few-shot CLIP classification challenge in which inference is performed jointly across a mini-batch of unlabeled query samples rather than treating each instance independently. We initially construct informative vision-text probability features leading to a classification problem on the unit simplex set. Inspired by Expectation-Maximization (EM) our optimization-based classifying objective models the data probability distribution for each class using a Dirichlet law. The minimization problem is then tackled with a novel block Majorization-Minimization algorithm which simultaneously estimates the distribution parameters and class assignments. Extensivenumerical experiments on 11 datasets underscore the benefits and efficacy of our batch inference approach. On zero-shot tasks with test batches of 75 samples our approach yields near 20% improvement in ImageNet accuracy over CLIP's zero-shot performance. Additionally we outperform state-of-the-art methods in the few-shot setting. Code is available at https://github.com/SegoleneMartin/transductive-CLIP.
****
Deep Single Image Camera Calibration by Heatmap Regression to Recover Fisheye Images Under Manhattan World Assumption
Nobuhiko Wakai, Satoshi Sato, Yasunori Ishii, Takayoshi Yamashita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11884-11894
A Manhattan world lying along cuboid buildings is useful for camera angle estimation. However accurate and robust angle estimation from fisheye images in the Manhattan world has remained an open challenge because general scene images tend to lack constraints such as lines arcs and vanishing points. To achieve higher accuracy and robustness we propose a learning-based calibration method that uses heatmap regression which is similar to pose estimation using keypoints to detect the directions of labeled image coordinates. Simultaneously our two estimators recover the rotation and remove fisheye distortion by remapping from a general scene image. Without considering vanishing-point constraints we find that additional points for learning-based methods can be defined. To compensate for the lack of vanishing points in images we introduce auxiliary diagonal points that have the optimal 3D arrangement of spatial uniformity. Extensive experiments demonstrated that our method outperforms conventional methods on large-scale datasets and with off-the-shelf cameras.
****
ID-Blau: Image Deblurring by Implicit Diffusion-based reBLurring AUgmentation
Jia-Hao Wu, Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Chia-Wen Lin, Yen-Yu Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25847-25856
Image deblurring aims to remove undesired blurs from an image captured in a dynamic scene. Much research has been dedicated to improving deblurring performance

through model architectural designs. However there is little work on data augmentation for image deblurring. Since continuous motion causes blurred artifacts during image exposure we aspire to develop a groundbreaking blur augmentation method to generate diverse blurred images by simulating motion trajectories in a continuous space. This paper proposes Implicit Diffusion-based reBLurring AUgmentation (ID-Blau) utilizing a sharp image paired with a controllable blur condition map to produce a corresponding blurred image. We parameterize the blur patterns of a blurred image with their orientations and magnitudes as a pixel-wise blur condition map to simulate motion trajectories and implicitly represent them in a continuous space. By sampling diverse blur conditions ID-Blau can generate various blurred images unseen in the training set. Experimental results demonstrate that ID-Blau can produce realistic blurred images for training and thus significantly improve performance for state-of-the-art deblurring models. The source code is available at https://github.com/plusgood-steven/ID-Blau.
****
LAENeRF: Local Appearance Editing for Neural Radiance Fields
Lukas Radl, Michael Steiner, Andreas Kurz, Markus Steinberger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4969-4978
Due to the omnipresence of Neural Radiance Fields (NeRFs) the interest towards editable implicit 3D representations has surged over the last years. However editing implicit or hybrid representations as used for NeRFs is difficult due to the entanglement of appearance and geometry encoded in the model parameters. Despite these challenges recent research has shown first promising steps towards photorealistic and non-photorealistic appearance edits. The main open issues of related work include limited interactivity a lack of support for local edits and large memory requirements rendering them less useful in practice. We address these limitations with LAENeRF a unified framework for photorealistic and non-photorealistic appearance editing of NeRFs. To tackle local editing we leverage a voxel grid as starting point for region selection. We learn a mapping from expected ray terminations to final output color which can optionally be supervised by a style loss resulting in a framework which can perform photorealistic and non-photorealistic appearance editing of selected regions. Relying on a single point per ray for our mapping we limit memory requirements and enable fast optimization. To guarantee interactivity we compose the output color using a set of learned modifiable base colors composed with additive layer mixing. Compared to concurrent work LAENeRF enables recoloring and stylization while keeping processing time low. Furthermore we demonstrate that our approach surpasses baseline methods both quantitatively and qualitatively.
****
CSTA: CNN-based Spatiotemporal Attention for Video Summarization
Jaewon Son, Jaehun Park, Kwangsu Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18847-18856
Video summarization aims to generate a concise representation of a video capturing its essential content and key moments while reducing its overall length. Although several methods employ attention mechanisms to handle long-term dependencies they often fail to capture the visual significance inherent in frames. To address this limitation we propose a CNN-based SpatioTemporal Attention (CSTA) method that stacks each feature of frames from a single video to form image-like frame representations and applies 2D CNN to these frame features. Our methodology relies on CNN to comprehend the inter and intra-frame relations and to find crucial attributes in videos by exploiting its ability to learn absolute positions within images. In contrast to previous work compromising efficiency by designing additional modules to focus on spatial importance CSTA requires minimal computational overhead as it uses CNN as a sliding window. Extensive experiments on two benchmark datasets (SumMe and TVSum) demonstrate that our proposed approach achieves state-of-the-art performance with fewer MACs compared to previous methods. Codes are available at https://github.com/thswodnjs3/CSTA.
****
Adversarial Score Distillation: When score distillation meets GAN

Min Wei, Jingkai Zhou, Junyao Sun, Xuesong Zhang; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8131-8141
Existing score distillation methods are sensitive to classifier-free guidance (C
FG) scale manifested as over-smoothness or instability at small CFG scales while
 over-saturation at large ones. To explain and analyze these issues we revisit t
he derivation of Score Distillation Sampling (SDS) and decipher existing score d
istillation with the Wasserstein Generative Adversarial Network (WGAN) paradigm.
 With the WGAN paradigm we find that existing score distillation either employs
a fixed sub-optimal discriminator or conducts incomplete discriminator optimizat
ion resulting in the scale-sensitive issue. We propose the Adversarial Score Dis
tillation (ASD) which maintains an optimizable discriminator and updates it usin
g the complete optimization objective. Experiments show that the proposed ASD pe
rforms favorably in 2D distillation and text-to-3D tasks against existing method
s. Furthermore to explore the generalization ability of our paradigm we extend A
SD to the image editing task which achieves competitive results. The project pag
e and code are at https://github.com/2y7c3/ASD
****
Decentralized Directed Collaboration for Personalized Federated Learning
Yingqi Liu, Yifan Shi, Qinglun Li, Baoyuan Wu, Xueqian Wang, Li Shen; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 23168-23178
Personalized Federated Learning (PFL) is proposed to find the greatest personali
zed models for each client. To avoid the central failure and communication bottl
eneck in the server-based FL we concentrate on the Decentralized Personalized Fe
derated Learning (DPFL) that performs distributed model training in a Peer-to-Pe
er (P2P) manner. Most personalized works in DPFL are based on undirected and sym
metric topologies however the data computation and communication resources heter
ogeneity result in large variances in the personalized models which lead the und
irected aggregation to suboptimal personalized performance and unguaranteed conv
ergence. To address these issues we propose a directed collaboration DPFL framew
ork by incorporating stochastic gradient push and partial model personalized cal
led Decentralized Federated Partial Gradient Push (DFedPGP). It personalizes the
 linear classifier in the modern deep model to customize the local solution and
learns a consensus representation in a fully decentralized manner. Clients only
share gradients with a subset of neighbors based on the directed and asymmetric
topologies which guarantees flexible choices for resource efficiency and better
convergence. Theoretically we show that the proposed DFedPGP achieves a superior
 convergence rate of O(1/?T) in the general non-convex setting and tighter conne
ctivity among clients will speed up the convergence. The proposed method achieve
s state-of-the-art (SOTA) accuracy in both data and computation heterogeneity sc
enarios demonstrating the efficiency of the directed collaboration and partial g
radient push.
****
Vector Graphics Generation via Mutually Impulsed Dual-domain Diffusion
Zhongyin Zhao, Ye Chen, Zhangli Hu, Xuanhong Chen, Bingbing Ni; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024,
pp. 4420-4428
Intelligent generation of vector graphics has very promising applications in the
 fields of advertising and logo design artistic painting animation production et
c. However current mainstream vector image generation methods lack the encoding
of image appearance information that is associated with the original vector repr
esentation and therefore lose valid supervision signal from the strong correlati
on between the discrete vector parameter (drawing instruction) sequence and the
target shape/structure of the corresponding pixel image. On the one hand the gen
eration process based on pure vector domain completely ignores the similarity me
asurement between shape parameter (and their combination) and the paired pixel i
mage appearance pattern; on the other hand two-stage methods (i.e. generation-an
d-vectorization) based on pixel diffusion followed by differentiable image-to-ve
ctor translation suffer from wrong error-correction signal caused by approximate
 gradients. To address the above issues we propose a novel generation framework

based on dual-domain (vector-pixel) diffusion with cross-modality impulse signals from each other. First in each diffusion step the current representation extracted from the other domain is used as a condition variable to constrain the subsequent sampling operation yielding shape-aware new parameterizations; second independent supervision signals from both domains avoid the gradient error accumulation problem caused by cross-domain representation conversion. Extensive experimental results on popular benchmarks including font and icon datasets demonstrate the great advantages of our proposed framework in terms of generated shape quality.
****
PEM: Prototype-based Efficient MaskFormer for Image Segmentation
Niccolò Cavagnero, Gabriele Rosi, Claudia Cuttano, Francesca Pistilli, Marco Ciccone, Giuseppe Averta, Fabio Cermelli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15804-15813
Recent transformer-based architectures have shown impressive results in the field of image segmentation. Thanks to their flexibility they obtain outstanding performance in multiple segmentation tasks such as semantic and panoptic under a single unified framework. To achieve such impressive performance these architectures employ intensive operations and require substantial computational resources which are often not available especially on edge devices. To fill this gap we propose Prototype-based Efficient MaskFormer (PEM) an efficient transformer-based architecture that can operate in multiple segmentation tasks. PEM proposes a novel prototype-based cross-attention which leverages the redundancy of visual features to restrict the computation and improve the efficiency without harming the performance. In addition PEM introduces an efficient multi-scale feature pyramid network capable of extracting features that have high semantic content in an efficient way thanks to the combination of deformable convolutions and context-based self-modulation. We benchmark the proposed PEM architecture on two tasks semantic and panoptic segmentation evaluated on two different datasets Cityscapes and ADE20K. PEM demonstrates outstanding performance on every task and dataset outperforming task-specific architectures while being comparable and even better than computationally expensive baselines. Code is available at https://github.com/NiccoloCavagnero/PEM.
****
Referring Expression Counting
Siyang Dai, Jun Liu, Ngai-Man Cheung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16985-16995
Existing counting tasks are limited to the class level which don't account for fine-grained details within the class. In real applications it often requires in-context or referring human input for counting target objects. Take urban analysis as an example fine-grained information such as traffic flow in different directions pedestrians and vehicles waiting or moving at different sides of the junction is more beneficial. Current settings of both class-specific and class-agnostic counting treat objects of the same class indifferently which pose limitations in real use cases. To this end we propose a new task named Referring Expression Counting (REC) which aims to count objects with different attributes within the same class. To evaluate the REC task we create a novel dataset named REC-8K which contains 8011 images and 17122 referring expressions. Experiments on REC-8K show that our proposed method achieves state-of-the-art performance compared with several text-based counting methods and an open-set object detection model. We also outperform prior models on the class agnostic counting (CAC) benchmark [36] for the zero-shot setting and perform on par with the few-shot methods. Code and dataset is available at https://github.com/sydai/referring-expression-counting.
****
ScoreHypo: Probabilistic Human Mesh Estimation with Hypothesis Scoring
Yuan Xu, Xiaoxuan Ma, Jiajun Su, Wentao Zhu, Yu Qiao, Yizhou Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 979-989
Monocular 3D human mesh estimation is an ill-posed problem characterized by inhe

rent ambiguity and occlusion. While recent probabilistic methods propose generating multiple solutions little attention is paid to obtaining high-quality estimates from them. To address this limitation we introduce ScoreHypo a versatile framework by first leverages our novel HypoNet to generate multiple hypotheses followed by employing a meticulously designed scorer ScoreNet to evaluate and select high-quality estimates. ScoreHypo formulates the estimation process as a reverse denoising process where HypoNet produces a diverse set of plausible estimates that effectively align with the image cues. Subsequently ScoreNet is employed to rigorously evaluate and rank these estimates based on their quality and finally identify superior ones. Experimental results demonstrate that HypoNet outperforms existing state-of-the-art probabilistic methods as a multi-hypothesis mesh estimator. Moreover the estimates selected by ScoreNet significantly outperform random generation or simple averaging. Notably the trained ScoreNet exhibits generalizability as it can effectively score existing methods and significantly reduce their errors by more than 15%. Code and models are available at https://xy02-05.github.io/ScoreHypo.
****
GES : Generalized Exponential Splatting for Efficient Radiance Field Rendering
Abdullah Hamdi, Luke Melas-Kyriazi, Jinjie Mai, Guocheng Qian, Ruoshi Liu, Carl Vondrick, Bernard Ghanem, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19812-19822
Advancements in 3D Gaussian Splatting have significantly accelerated 3D reconstruction and generation. However it may require a large number of Gaussians which creates a substantial memory footprint. This paper introduces GES (Generalized Exponential Splatting) a novel representation that employs Generalized Exponential Function (GEF) to model 3D scenes requiring far fewer particles to represent a scene and thus significantly outperforming Gaussian Splatting methods in efficiency with a plug-and-play replacement ability for Gaussian-based utilities. GES is validated theoretically and empirically in both principled 1D setup and realistic 3D scenes. It is shown to represent signals with sharp edges more accurately which are typically challenging for Gaussians due to their inherent low-pass characteristics. Our empirical analysis demonstrates that GEF outperforms Gaussians in fitting natural-occurring signals (E.g. squares triangles parabolic signals) thereby reducing the need for extensive splitting operations that increase the memory footprint of Gaussian Splatting. With the aid of a frequency-modulated loss GES achieves competitive performance in novel-view synthesis benchmarks while requiring less than half the memory storage of Gaussian Splatting and increasing the rendering speed by up to 39%. The code is available on the project website https://abdullahamdi.com/ges .
****
Learning to Predict Activity Progress by Self-Supervised Video Alignment
Gerard Donahue, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18667-18677
In this paper we tackle the problem of self-supervised video alignment and activity progress prediction using in-the-wild videos. Our proposed self-supervised representation learning method carefully addresses different action orderings redundant actions and background frames to generate improved video representations compared to previous methods. Our model generalizes temporal cycle-consistency learning to allow for more flexibility in determining cycle-consistent neighbors. More specifically to handle repeated actions we propose a multi-neighbor cycle consistency and a multi-cycle-back regression loss by finding multiple soft nearest neighbors using a Gaussian Mixture Model. To handle background and redundant frames we introduce a context-dependent drop function in our framework discouraging the alignment of droppable frames. On the other hand to learn from videos of multiple activities jointly we propose a multi-head crosstask network allowing us to embed a video and estimate progress without knowing its activity label. Experiments on multiple datasets show that our method outperforms the state-of-the-art for video alignment and progress prediction.
****
VicTR: Video-conditioned Text Representations for Activity Recognition

Kumara Kahatapitiya, Anurag Arnab, Arsha Nagrani, Michael S. Ryoo; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 18547-18558
Vision-Language models (VLMs) have excelled in the image-domain--- especially in
 zero-shot settings--- thanks to the availability of vast pretraining data (i.e.
 paired image-text samples). However for videos such paired data is not as abund
ant. Therefore video-VLMs are usually designed by adapting pretrained image-VLMs
 to the video-domain instead of training from scratch. All such recipes rely on
augmenting visual embeddings with temporal information (i.e. image --> video) of
ten keeping text embeddings unchanged or even being discarded. In this paper we
argue the contrary that better video-VLMs can be designed by focusing more on au
gmenting text rather than visual information. More specifically we introduce Vid
eo-conditioned Text Representations (VicTR): a form of text embeddings optimized
 w.r.t. visual embeddings creating a more-flexible contrastive latent space. Our
 model can further make use of freely-available semantic information in the form
 of visually-grounded auxiliary text (e.g. object or scene information). We eval
uate our model on few-shot zero-shot (HMDB-51 UCF-101) short-form (Kinetics-400)
 and long-form (Charades) activity recognition benchmarks showing strong perform
ance among video-VLMs.
****
Label-Efficient Group Robustness via Out-of-Distribution Concept Curation
Yiwei Yang, Anthony Z. Liu, Robert Wolfe, Aylin Caliskan, Bill Howe; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 12426-12434
Deep neural networks are prone to capture correlations between spurious attribut
es and class labels leading to low accuracy on some combinations of class labels
 and spurious attribute values. When a spurious attribute represents a protected
 class these low-accuracy groups can manifest discriminatory bias. Existing meth
ods attempting to improve worst-group accuracy assume the training data validati
on data or both are reliably labeled by the spurious attribute. But a model may
be perceived to be biased towards a concept that is not represented by pre-exist
ing labels on the training data. In these situations the spurious attribute must
 be defined with external information. We propose Concept Correction a framework
 that represents a concept as a curated set of images from any source then label
s each training sample by its similarity to the concept set to control spurious
correlations. For example concept sets representing gender can be used to measur
e and control gender bias even without explicit labels. We demonstrate and evalu
ate an instance of the framework as Concept DRO which uses concept sets to estim
ate group labels then uses these labels to train with a state of the art distrib
utively robust optimization objective. We show that Concept DRO outperforms exis
ting methods that do not require labels of spurious attributes by up to 33.1% on
 three image classification datasets and is competitive with the best methods th
at assume access to labels. We consider how the size and quality of the concept
set influences performance and find that even smaller manually curated sets of n
oisy AI-generated images are effective at controlling spurious correlations sugg
esting that high-quality reusable concept sets are easy to create and effective
in reducing bias.
****
MMCert: Provable Defense against Adversarial Attacks to Multi-modal Models
Yanting Wang, Hongye Fu, Wei Zou, Jinyuan Jia; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24655-24664
Different from a unimodal model whose input is from a single modality the input
(called multi-modal input) of a multi-modal model is from multiple modalities su
ch as image 3D points audio text etc. Similar to unimodal models many existing s
tudies show that a multi-modal model is also vulnerable to adversarial perturbat
ion where an attacker could add small perturbation to all modalities of a multi-
modal input such that the multi-modal model makes incorrect predictions for it.
Existing certified defenses are mostly designed for unimodal models which achiev
e sub-optimal certified robustness guarantees when extended to multi-modal model
s as shown in our experimental results. In our work we propose MMCert the first

certified defense against adversarial attacks to a multi-modal model. We derive a lower bound on the performance of our MMCert under arbitrary adversarial attacks with bounded perturbations to both modalities (e.g. in the context of auto-driving we bound the number of changed pixels in both RGB image and depth image). We evaluate our MMCert using two benchmark datasets: one for the multi-modal road segmentation task and the other for the multi-modal emotion recognition task. Moreover we compare our MMCert with a state-of-the-art certified defense extended from unimodal models. Our experimental results show that our MMCert outperforms the baseline.
****
3DToonify: Creating Your High-Fidelity 3D Stylized Avatar Easily from 2D Portrait Images
Yifang Men, Hanxi Liu, Yuan Yao, Miaomiao Cui, Xuansong Xie, Zhouhui Lian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10127-10137
Visual content creation has aroused a surge of interest given its applications in mobile photography and AR/VR. Portrait style transfer and 3D recovery from monocular images as two representative tasks have so far evolved independently. In this paper we make a connection between the two and tackle the challenging task of 3D portrait stylization - modeling high-fidelity 3D stylized avatars from captured 2D portrait images. However naively combining the techniques from the two isolated areas may suffer from either inadequate stylization or absence of 3D assets. To this end we propose 3DToonify a new framework that introduces a progressive training scheme to achieve 3D style adaption on spatial neural representation (SNR). SNR is constructed with implicit fields and they are dynamically optimized by the progressive training scheme which consists of three stages: guided prior learning deformable geometry adaption and explicit texture adaption. In this way stylized geometry and texture are learned in SNR in an explicit and structured way with only a single stylized exemplar needed. Moreover our method obtains style-adaptive underlying structures (i.e. deformable geometry and exaggerated texture) and view-consistent stylized avatar rendering from arbitrary novel viewpoints. Both qualitative and quantitative experiments have been conducted to demonstrate the effectiveness and superiority of our method for automatically generating exemplar-guided 3D stylized avatars.
****
NAYER: Noisy Layer Data Generation for Efficient and Effective Data-free Knowledge Distillation
Minh-Tuan Tran, Trung Le, Xuan-May Le, Mehrtash Harandi, Quan Hung Tran, Dinh Phung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23860-23869
Data-Free Knowledge Distillation (DFKD) has made significant recent strides by transferring knowledge from a teacher neural network to a student neural network without accessing the original data. Nonetheless existing approaches encounter a significant challenge when attempting to generate samples from random noise inputs which inherently lack meaningful information. Consequently these models struggle to effectively map this noise to the ground-truth sample distribution resulting in prolonging training times and low-quality outputs. In this paper we propose a novel Noisy Layer Generation method (NAYER) which relocates the random source from the input to a noisy layer and utilizes the meaningful constant label-text embedding (LTE) as the input. LTE is generated by using the language model once and then it is stored in memory for all subsequent training processes. The significance of LTE lies in its ability to contain substantial meaningful inter-class information enabling the generation of high-quality samples with only a few training steps. Simultaneously the noisy layer plays a key role in addressing the issue of diversity in sample generation by preventing the model from overemphasizing the constrained label information. By reinitializing the noisy layer in each iteration we aim to facilitate the generation of diverse samples while still retaining the method's efficiency thanks to the ease of learning provided by LTE. Experiments carried out on multiple datasets demonstrate that our NAYER not only outperforms the state-of-the-art methods but also achieves speeds 5 to 15 t

imes faster than previous approaches. The code is available at https://github.co m/tmtuan1307/nayer.
****
OmniVec2 - A Novel Transformer based Network for Large Scale Multimodal and Mult itask Learning
Siddharth Srivastava, Gaurav Sharma; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2024, pp. 27412-27424
We present a novel multimodal multitask network and associated training algorith m. The method is capable of ingesting data from approximately 12 different modal ities namely image video audio text depth point cloud time series tabular graph X-ray infrared IMU and hyperspectral. The proposed approach utilizes modality sp ecialized tokenizers a shared transformer architecture and cross-attention mecha nisms to project the data from different modalities into a unified embedding spa ce. It addresses multimodal and multitask scenarios by incorporating modality-sp ecific task heads for different tasks in respective modalities. We propose a nov el pretraining strategy with iterative modality switching to initialize the netw ork and a training algorithm which trades off fully joint training over all moda lities with training on pairs of modalities at a time. We provide comprehensive evaluation across 25 datasets from 12 modalities and show state of the art perfo rmances demonstrating the effectiveness of the proposed architecture pretraining strategy and adapted multitask training.
****
Investigating Compositional Challenges in Vision-Language Models for Visual Grou nding
Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, Liang Wang; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2024, pp. 14141-14151
Pre-trained vision-language models (VLMs) have achieved high performance on vari ous downstream tasks which have been widely used for visual grounding tasks in a weakly supervised manner. However despite the performance gains contributed by large vision and language pre-training we find that state-of-the-art VLMs strugg le with compositional reasoning on grounding tasks. To demonstrate this we propo se Attribute Relation and Priority grounding (ARPGrounding) benchmark to test VL Ms' compositional reasoning ability on visual grounding tasks. ARPGrounding cont ains 11425 samples and evaluates the compositional understanding of VLMs in thre e dimensions: 1) attribute denoting comprehension of objects' properties; 2) rel ation indicating an understanding of relation between objects; 3) priority refle cting an awareness of the part of speech associated with nouns. Using the ARPGro unding benchmark we evaluate several mainstream VLMs. We empirically find that t hese models perform quite well on conventional visual grounding datasets achievi ng performance comparable to or surpassing state-of-the-art methods but showing strong deficiencies in compositional reasoning. Furthermore we propose a composi tion-aware fine-tuning pipeline demonstrating the potential to leverage cost-eff ective image-text annotations for enhancing the compositional understanding of V LMs in grounding tasks.
****
6D-Diff: A Keypoint Diffusion Framework for 6D Object Pose Estimation
Li Xu, Haoxuan Qu, Yujun Cai, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9676-9686
Estimating the 6D object pose from a single RGB image often involves noise and i ndeterminacy due to challenges such as occlusions and cluttered backgrounds. Mea nwhile diffusion models have shown appealing performance in generating high-qual ity images from random noise with high indeterminacy through step-by-step denois ing. Inspired by their denoising capability we propose a novel diffusion-based f ramework (6D-Diff) to handle the noise and indeterminacy in object pose estimati on for better performance. In our framework to establish accurate 2D-3D correspo ndence we formulate 2D keypoints detection as a reverse diffusion (denoising) pr ocess. To facilitate such a denoising process we design a Mixture-of-Cauchy-base d forward diffusion process and condition the reverse process on the object appe arance features. Extensive experiments on the LM-O and YCB-V datasets demonstrat

e the effectiveness of our framework.
****
Generative Region-Language Pretraining for Open-Ended Object Detection
Chuang Lin, Yi Jiang, Lizhen Qu, Zehuan Yuan, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13958-13968
In recent research significant attention has been devoted to the open-vocabulary object detection task aiming to generalize beyond the limited number of classes labeled during training and detect objects described by arbitrary category names at inference. Compared with conventional object detection open vocabulary object detection largely extends the object detection categories. However it relies on calculating the similarity between image regions and a set of arbitrary category names with a pretrained vision-and-language model. This implies that despite its open-set nature the task still needs the predefined object categories during the inference stage. This raises the question: What if we do not have exact knowledge of object categories during inference? In this paper we call such a new setting as generative open-ended object detection which is a more general and practical problem. To address it we formulate object detection as a generative problem and propose a simple framework named GenerateU which can detect dense objects and generate their names in a free-form way. Particularly we employ Deformable DETR as a region proposal generator with a language model translating visual regions to object names. To assess the free-form object detection task we introduce an evaluation method designed to quantitatively measure the performance of generative outcomes. Extensive experiments demonstrate strong zero-shot detection performance of our GenerateU. For example on the LVIS dataset our GenerateU achieves comparable results to the open-vocabulary object detection method GLIP even though the category names are not seen by GenerateU during inference. Code is available at: https://github.com/FoundationVision/GenerateU.
****
Enhancing Post-training Quantization Calibration through Contrastive Learning
Yuzhang Shang, Gaowen Liu, Ramana Rao Kompella, Yan Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15921-15930
Post-training quantization (PTQ) converts a pre-trained full-precision (FP) model into a quantized model in a training-free manner. Determining suitable quantization parameters such as scaling factors and weight rounding is the primary strategy for mitigating the impact of quantization noise (calibration) and restoring the performance of the quantized models. However the existing activation calibration methods have never considered information degradation between pre- (FP) and post-quantized activations. In this study we introduce a well-defined distributional metric from information theory mutual information into PTQ calibration. We aim to calibrate the quantized activations by maximizing the mutual information between the pre- and post-quantized activations. To realize this goal we establish a contrastive learning (CL) framework for the PTQ calibration where the quantization parameters are optimized through a self-supervised proxy task. Specifically by leveraging CL during the PTQ process we can benefit from pulling the positive pairs of quantized and FP activations collected from the same input samples while pushing negative pairs from different samples. Thanks to the ingeniously designed critic function we avoid the unwanted but often-encountered collision solution in CL especially in calibration scenarios where the amount of calibration data is limited. Additionally we provide a theoretical guarantee that minimizing our designed loss is equivalent to maximizing the desired mutual information. Consequently the quantized activations retain more information which ultimately enhances the performance of the quantized network. Experimental results show that our method can effectively serve as an add-on module to existing SoTA PTQ methods.
****
Efficient Model Stealing Defense with Noise Transition Matrix
Dong-Dong Wu, Chilin Fu, Weichang Wu, Wenwen Xia, Xiaolu Zhang, Jun Zhou, Min-Ling Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), 2024, pp. 24305-24315
With the escalating complexity and investment cost of training deep neural netwo
rks safeguarding them from unauthorized usage and intellectual property theft ha
s become imperative. Especially the rampant misuse of prediction APIs to replica
te models without access to the original data or architecture poses grave securi
ty threats. Diverse defense strategies have emerged to address these vulnerabili
ties yet these defenses either incur heavy inference overheads or assume idealiz
ed attack scenarios. To address these challenges we revisit the utilization of n
oise transition matrix as an efficient perturbation technique which injects nois
e into predicted posteriors in a linear manner and integrates seamlessly into ex
isting systems with minimal overhead for model stealing defense. Provably with s
uch perturbed posteriors the attacker's cloning process degrades into learning f
rom noisy data. Toward optimizing the noise transition matrix we proposed a nove
l bi-level optimization training framework which performs fidelity on the victim
 model while the surrogate model adversarially. Comprehensive experimental resul
ts demonstrate that our method effectively thwarts model stealing attacks and ac
hieves minimal utility trade-offs outperforming existing state-of-the-art defens
es.
****
MeshPose: Unifying DensePose and 3D Body Mesh Reconstruction
Eric-Tuan Le, Antonis Kakolyris, Petros Koutras, Himmy Tam, Efstratios Skordos,
George Papandreou, Riza Alp Güler, Iasonas Kokkinos; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2405-24
14
DensePose provides a pixel-accurate association of images with 3D mesh coordinat
es but does not provide a 3D mesh while Human Mesh Reconstruction (HMR) systems
have high 2D reprojection error as measured by DensePose localization metrics. I
n this work we introduce MeshPose to jointly tackle DensePose and HMR. For this
we first introduce new losses that allow us to use weak DensePose supervision to
 accurately localize in 2D a subset of the mesh vertices ('VertexPose'). We then
 lift these vertices to 3D yielding a low-poly body mesh ('MeshPose'). Our syste
m is trained in an end-to-end manner and is the first HMR method to attain compe
titive DensePose accuracy while also being lightweight and amenable to efficient
 inference making it suitable for real-time AR applications.
****
Unsupervised Salient Instance Detection
Xin Tian, Ke Xu, Rynson Lau; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2024, pp. 2702-2712
The significant amount of manual efforts in annotating pixel-level labels has tr
iggered the advancement of unsupervised saliency learning. However without super
vision signals state-of-the-art methods can only infer region-level saliency. In
 this paper we propose to explore the unsupervised salient instance detection (U
SID) problem for a more fine-grained visual understanding. Our key observation i
s that self-supervised transformer features may exhibit local similarities as we
ll as different levels of contrast to other regions which provide informative cu
es to identify salient instances. Hence we propose SCoCo a novel network that mo
dels saliency coherence and contrast for USID. SCoCo includes two novel modules:
 (1) a global background adaptation (GBA) module with a scene-level contrastive
loss to extract salient regions from the scene by searching the adaptive "salien
cy threshold" in the self-supervised transformer features and (2) a locality-awa
re similarity (LAS) module with an instance-level contrastive loss to group sali
ent regions into instances by modeling the in-region saliency coherence and cros
s-region saliency contrasts. Extensive experiments show that SCoCo outperforms s
tate-of-the-art weakly-supervised SID methods and carefully designed unsupervise
d baselines and has comparable performances to fully-supervised SID methods.
****
Enhancing Visual Document Understanding with Contrastive Learning in Large Visua
l-Language Models
Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong
Liu, Deqiang Jiang, Xing Sun; Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR), 2024, pp. 15546-15555
Recently the advent of Large Visual-Language Models (LVLMs) has received increasing attention across various domains particularly in the field of visual document understanding (VDU). Different from conventional vision-language tasks VDU is specifically concerned with text-rich scenarios containing abundant document elements. Nevertheless the importance of fine-grained features remains largely unexplored within the community of LVLMs leading to suboptimal performance in text-rich scenarios. In this paper we abbreviate it as the fine-grained feature collapse issue. With the aim of filling this gap we propose a contrastive learning framework termed Document Object COntrastive learning (DoCo) specifically tailored for the downstream tasks of VDU. DoCo leverages an auxiliary multimodal encoder to obtain the features of document objects and align them to the visual features generated by the vision encoder of LVLM which enhances visual representation in text-rich scenarios. It can represent that the contrastive learning between the visual holistic representations and the multimodal fine-grained features of document objects can assist the vision encoder in acquiring more effective visual cues thereby enhancing the comprehension of text-rich documents in LVLMs. We also demonstrate that the proposed DoCo serves as a plug-and-play pre-training method which can be employed in the pre-training of various LVLMs without inducing any increase in computational complexity during the inference process. Extensive experimental results on multiple benchmarks of VDU reveal that LVLMs equipped with our proposed DoCo can achieve superior performance and mitigate the gap between VDU and generic vision-language tasks.
****
Move Anything with Layered Scene Diffusion
Jiawei Ren, Mengmeng Xu, Jui-Chieh Wu, Ziwei Liu, Tao Xiang, Antoine Toisoul; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6380-6389
Diffusion models generate images with an unprecedented level of quality but how can we freely rearrange image layouts? Recent works generate controllable scenes via learning spatially disentangled latent codes but these methods do not apply to diffusion models due to their fixed forward process. In this work we propose SceneDiffusion to optimize a layered scene representation during the diffusion sampling process. Our key insight is that spatial disentanglement can be obtained by jointly denoising scene renderings at different spatial layouts. Our generated scenes support a wide range of spatial editing operations including moving resizing cloning and layer-wise appearance editing operations including object restyling and replacing. Moreover a scene can be generated conditioned on a reference image thus enabling object moving for in-the-wild images. Notably this approach is training-free compatible with general text-to-image diffusion models and responsive in less than a second.
****
GS-SLAM: Dense Visual SLAM with 3D Gaussian Splatting
Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, Xuelong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19595-19604
In this paper we introduce GS-SLAM that first utilizes 3D Gaussian representation in the Simultaneous Localization and Mapping (SLAM) system. It facilitates a better balance between efficiency and accuracy. Compared to recent SLAM methods employing neural implicit representations our method utilizes a real-time differentiable splatting rendering pipeline that offers significant speedup to map optimization and RGB-D rendering. Specifically we propose an adaptive expansion strategy that adds new or deletes noisy 3D Gaussians in order to efficiently reconstruct new observed scene geometry and improve the mapping of previously observed areas. This strategy is essential to extend 3D Gaussian representation to reconstruct the whole scene rather than synthesize a static object in existing methods. Moreover in the pose tracking process an effective coarse-to-fine technique is designed to select reliable 3D Gaussian representations to optimize camera pose resulting in runtime reduction and robust estimation. Our method achieves competitive performance compared with existing state-of-the-art real-time methods on

the Replica TUM-RGBD datasets. Project page: \href https://gs-slam.github.io/ h
ttps://gs-slam.github.io/ .
****
Scaffold-GS: Structured 3D Gaussians for View-Adaptive Rendering
Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, Bo Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20654-20664
Neural rendering methods have significantly advanced photo-realistic 3D scene rendering in various academic and industrial applications. The recent 3D Gaussian Splatting method has achieved the state-of-the-art rendering quality and speed combining the benefits of both primitive-based representations and volumetric representations. However it often leads to heavily redundant Gaussians that try to fit every training view neglecting the underlying scene geometry. Consequently the resulting model becomes less robust to significant view changes texture-less area and lighting effects. We introduce Scaffold-GS which uses anchor points to distribute local 3D Gaussians and predicts their attributes on-the-fly based on viewing direction and distance within the view frustum. Anchor growing and pruning strategies are developed based on the importance of neural Gaussians to reliably improve the scene coverage. We show that our method effectively reduces redundant Gaussians while delivering high-quality rendering. We also demonstrates an enhanced capability to accommodate scenes with varying levels-of-detail and view-dependent observations without sacrificing the rendering speed. Project page: https://city-super.github.io/scaffold-gs.
****
Data Valuation and Detections in Federated Learning
Wenqian Li, Shuran Fu, Fengrui Zhang, Yan Pang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12027-12036
Federated Learning (FL) enables collaborative model training while preserving the privacy of raw data. A challenge in this framework is the fair and efficient valuation of data which is crucial for incentivizing clients to contribute high-quality data in the FL task. In scenarios involving numerous data clients within FL it is often the case that only a subset of clients and datasets are pertinent to a specific learning task while others might have either a negative or negligible impact on the model training process. This paper introduces a novel privacy-preserving method for evaluating client contributions and selecting relevant datasets without a pre-specified training algorithm in an FL task. Our proposed approach \texttt FedBary utilizes Wasserstein distance within the federated context offering a new solution for data valuation in the FL framework. This method ensures transparent data valuation and efficient computation of the Wasserstein barycenter and reduces the dependence on validation datasets. Through extensive empirical experiments and theoretical analyses we demonstrate the advantages of this data valuation method as a promising avenue for FL research.
****
Classes Are Not Equal: An Empirical Study on Image Recognition Fairness
Jiequan Cui, Beier Zhu, Xin Wen, Xiaojuan Qi, Bei Yu, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23283-23292
In this paper we present an empirical study on image recognition unfairness i.e. extreme class accuracy disparity on balanced data like ImageNet. We demonstrate that classes are not equal and unfairness is prevalent for image classification models across various datasets network architectures and model capacities. Moreover several intriguing properties of fairness are identified. First the unfairness lies in problematic representation rather than classifier bias distinguished from long-tailed recognition. Second with the proposed concept of Model Prediction Bias we investigate the origins of problematic representation during training optimization. Our findings reveal that models tend to exhibit greater prediction biases for classes that are more challenging to recognize. It means that more other classes will be confused with harder classes. Then the False Positives (FPs) will dominate the learning in optimization thus leading to their poor accuracy. Further we conclude that data augmentation and representation learning algor

ithms improve overall performance by promoting fairness to some degree in image classification.
****
Human Gaussian Splatting: Real-time Rendering of Animatable Avatars
Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, Eduardo Pérez-Pellitero; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 788-798
This work addresses the problem of real-time rendering of photorealistic human body avatars learned from multi-view videos. While the classical approaches to model and render virtual humans generally use a textured mesh recent research has developed neural body representations that achieve impressive visual quality. However these models are difficult to render in real-time and their quality degrades when the character is animated with body poses different than the training observations. We propose an animatable human model based on 3D Gaussian Splatting that has recently emerged as a very efficient alternative to neural radiance fields. The body is represented by a set of gaussian primitives in a canonical space which is deformed with a coarse to fine approach that combines forward skinning and local non-rigid refinement. We describe how to learn our Human Gaussian Splatting (HuGS) model in an end-to-end fashion from multi-view observations and evaluate it against the state-of-the-art approaches for novel pose synthesis of clothed body. Our method achieves 1.5 dB PSNR improvement over the state-of-the-art on THuman4 dataset while being able to render in real-time (?80 fps for 512x512 resolution).
****
Multi-Scale 3D Gaussian Splatting for Anti-Aliased Rendering
Zhiwen Yan, Weng Fei Low, Yu Chen, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20923-20931
3D Gaussians have recently emerged as a highly efficient representation for 3D reconstruction and rendering. Despite its high rendering quality and speed at high resolutions they both deteriorate drastically when rendered at lower resolutions or from far away camera position. During low resolution or far away rendering the pixel size of the image can fall below the Nyquist frequency compared to the screen size of each splatted 3D Gaussian and leads to aliasing effect. The rendering is also drastically slowed down by the sequential alpha blending of more splatted Gaussians per pixel. To address these issues we propose a multi-scale 3D Gaussian splatting algorithm which maintains Gaussians at different scales to represent the same scene. Higher-resolution images are rendered with more small Gaussians and lower-resolution images are rendered with fewer larger Gaussians. With similar training time our algorithm can achieve 13%-66% PSNR and 160%-2400% rendering speed improvement at 4x-128x scale rendering on Mip-NeRF360 dataset compared to the single scale 3D Gaussian splatting.
****
A Bayesian Approach to OOD Robustness in Image Classification
Prakhar Kaushik, Adam Kortylewski, Alan Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22988-22997
An important and unsolved problem in computer vision is to ensure that the algorithms are robust to changes in image domains. We address this problem in the scenario where we have access to images from the target domains but no annotations. Motivated by the challenges of the OOD-CV benchmark where we encounter real world Out-of-Domain (OOD) nuisances and occlusion we introduce a novel Bayesian approach to OOD robustness for object classification. Our work extends Compositional Neural Networks (CompNets) which have been shown to be robust to occlusion but degrade badly when tested on OOD data. We exploit the fact that CompNets contain a generative head defined over feature vectors represented by von Mises-Fisher (vMF) kernels which correspond roughly to object parts and can be learned without supervision. We obverse that some vMF kernels are similar between different domains while others are not. This enables us to learn a transitional dictionary of vMF kernels that are intermediate between the source and target domains and train the generative model on this dictionary using the annotations on the source domain followed by iterative refinement. This approach termed Unsupervised Gene

rative Transition (UGT) performs very well in OOD scenarios even when occlusion is present. UGT is evaluated on different OOD benchmarks including the OOD-CV dataset several popular datasets (e.g. ImageNet-C artificial image corruptions (including adding occluders) and synthetic-to-real domain transfer and does well in all scenarios outperforming SOTA alternatives.
****
Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action
Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, Aniruddha Kembhavi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26439-26455
We present Unified-IO 2 a multimodal and multi-skill unified model capable of following novel instructions. Unified-IO 2 can use text images audio and/or videos as input and can generate text image or audio outputs which is accomplished in a unified way by tokenizing these different inputs and outputs into a shared semantic space that can then be processed by a single encoder-decoder transformer model. Unified-IO 2 is trained from scratch on a custom-built multimodal pre-training corpus and then learns an expansive set of skills through fine-tuning on over 120 datasets including datasets for segmentation object detection image editing audio localization video tracking embodied AI and 3D detection. To facilitate instruction-following we add prompts and other data augmentations to these tasks to allow Unified-IO 2 to generalize these skills to new tasks zero-shot. Unified-IO 2 is the first model to be trained on such a diverse and wide-reaching set of skills and unify three separate generation capabilities. Unified-IO 2 achieves state-of-the-art performance on the multi-task GRIT benchmark and achieves strong results on 30 diverse datasets including SEED-Bench image and video understanding TIFA image generation VQA 2.0 ScienceQA VIMA robotic manipulation VGG-Sound and Kinetics-Sounds and can perform unseen tasks and generate free-form responses. We release our model and code to facilitate future work.
****
Joint Reconstruction of 3D Human and Object via Contact-Based Refinement Transformer
Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10218-10227
Human-object contact serves as a strong cue to understand how humans physically interact with objects. Nevertheless it is not widely explored to utilize human-object contact information for the joint reconstruction of 3D human and object from a single image. In this work we present a novel joint 3D human-object reconstruction method (CONTHO) that effectively exploits contact information between humans and objects. There are two core designs in our system: 1) 3D-guided contact estimation and 2) contact-based 3D human and object refinement. First for accurate human-object contact estimation CONTHO initially reconstructs 3D humans and objects and utilizes them as explicit 3D guidance for contact estimation. Second to refine the initial reconstructions of 3D human and object we propose a novel contact-based refinement Transformer that effectively aggregates human features and object features based on the estimated human-object contact. The proposed contact-based refinement prevents the learning of erroneous correlation between human and object which enables accurate 3D reconstruction. As a result our CONTHO achieves state-of-the-art performance in both human-object contact estimation and joint reconstruction of 3D human and object. The codes are available in https://github.com/dqj5182/CONTHO_RELEASE.
****
TIM: A Time Interval Machine for Audio-Visual Action Recognition
Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, Dima Damen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18153-18163
Diverse actions give rise to rich audio-visual signals in long videos. Recent works showcase that the two modalities of audio and video exhibit different temporal extents of events and distinct labels. We address the interplay between the t

wo modalities in long videos by explicitly modelling the temporal extents of aud io and visual events. We propose the Time Interval Machine (TIM) where a modalit y-specific time interval poses as a query to a transformer encoder that ingests a long video input. The encoder then attends to the specified interval as well a s the surrounding context in both modalities in order to recognise the ongoing a ction. We test TIM on three long audio-visual video datasets: EPIC-KITCHENS Perc eption Test and AVE reporting state-of-the-art (SOTA) for recognition. On EPIC-K ITCHENS we beat previous SOTA that utilises LLMs and significantly larger pre-tr aining by 2.9% top-1 action recognition accuracy. Additionally we show that TIM can be adapted for action detection using dense multi-scale interval queries out performing SOTA on EPIC-KITCHENS-100 for most metrics and showing strong perform ance on the Perception Test. Our ablations show the critical role of integrating the two modalities and modelling their time intervals in achieving this perform ance. Code and models at: https://github.com/JacobChalk/TIM.
****
The Devil is in the Details: StyleFeatureEditor for Detail-Rich StyleGAN Inversi on and High Quality Image Editing
Denis Bobkov, Vadim Titov, Aibek Alanov, Dmitry Vetrov; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9337 -9346
The task of manipulating real image attributes through StyleGAN inversion has be en extensively researched. This process involves searching latent variables from a well-trained StyleGAN generator that can synthesize a real image modifying th ese latent variables and then synthesizing an image with the desired edits. A ba lance must be struck between the quality of the reconstruction and the ability t o edit. Earlier studies utilized the low-dimensional W-space for latent search w hich facilitated effective editing but struggled with reconstructing intricate d etails. More recent research has turned to the high-dimensional feature space F which successfully inverses the input image but loses much of the detail during editing. In this paper we introduce StyleFeatureEditor -- a novel method that en ables editing in both w-latents and F-latents. This technique not only allows fo r the reconstruction of finer image details but also ensures their preservation during editing. We also present a new training pipeline specifically designed to train our model to accurately edit F-latents. Our method is compared with state -of-the-art encoding approaches demonstrating that our model excels in terms of reconstruction quality and is capable of editing even challenging out-of-domain examples.
****
Unbiased Estimator for Distorted Conics in Camera Calibration
Chaehyeon Song, Jaeho Shin, Myung-Hwan Jeon, Jongwoo Lim, Ayoung Kim; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 373-381
In the literature points and conics have been major features for camera geometri c calibration. Although conics are more informative features than points the los s of the conic property under distortion has critically limited the utility of c onic features in camera calibration. Many existing approaches addressed conic-ba sed calibration by ignoring distortion or introducing 3D spherical targets to ci rcumvent this limitation. In this paper we present a novel formulation for conic -based calibration using moments. Our derivation is based on the mathematical fi nding that the first moment can be estimated without bias even under distortion. This allows us to track moment changes during projection and distortion ensurin g the preservation of the first moment of the distorted conic. With an unbiased estimator the circular patterns can be accurately detected at the sub-pixel leve l and can now be fully exploited for an entire calibration pipeline resulting in significantly improved calibration. The entire code is readily available from h ttps://github.com/ChaehyeonSong/discocal.
****
MultiPhys: Multi-Person Physics-aware 3D Motion Estimation
Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Sh en, Jordi Sanchez-Riera, Francesc Moreno-Noguer, Leonidas Guibas; Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2331-2340
We introduce MultiPhys a method designed for recovering multi-person motion from monocular videos. Our focus lies in capturing coherent spatial placement between pairs of individuals across varying degrees of engagement. MultiPhys being physically aware exhibits robustness to jittering and occlusions and effectively eliminates penetration issues between the two individuals. We devise a pipeline in which the motion estimated by a kinematic-based method is fed into a physics simulator in an autoregressive manner. We introduce distinct components that enable our model to harness the simulator's properties without compromising the accuracy of the kinematic estimates. This results in final motion estimates that are both kinematically coherent and physically compliant. Extensive evaluations on three challenging datasets characterized by substantial inter-person interaction show that our method significantly reduces errors associated with penetration and foot skating while performing competitively with the state-of-the-art on motion accuracy and smoothness.
****
Multi-Level Neural Scene Graphs for Dynamic Urban Environments
Tobias Fischer, Lorenzo Porzi, Samuel Rota Bulo, Marc Pollefeys, Peter Kontschieder; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21125-21135
We estimate the radiance field of large-scale dynamic areas from multiple vehicle captures under varying environmental conditions. Previous works in this domain are either restricted to static environments do not scale to more than a single short video or struggle to separately represent dynamic object instances. To this end we present a novel decomposable radiance field approach for dynamic urban environments. We propose a multi-level neural scene graph representation that scales to thousands of images from dozens of sequences with hundreds of fast-moving objects. To enable efficient training and rendering of our representation we develop a fast composite ray sampling and rendering scheme. To test our approach in urban driving scenarios we introduce a new novel view synthesis benchmark. We show that our approach outperforms prior art by a significant margin on both established and our proposed benchmark while being faster in training and rendering.
****
Would Deep Generative Models Amplify Bias in Future Models?
Tianwei Chen, Yusuke Hirota, Mayu Otani, Noa Garcia, Yuta Nakashima; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10833-10843
We investigate the impact of deep generative models on potential social biases in upcoming computer vision models. As the internet witnesses an increasing influx of AI-generated images concerns arise regarding inherent biases that may accompany them potentially leading to the dissemination of harmful content. This paper explores whether a detrimental feedback loop resulting in bias amplification would occur if generated images were used as the training data for future models. We conduct simulations by progressively substituting original images in COCO and CC3M datasets with images generated through Stable Diffusion. The modified datasets are used to train OpenCLIP and image captioning models which we evaluate in terms of quality and bias. Contrary to expectations our findings indicate that introducing generated images during training does not uniformly amplify bias. Instead instances of bias mitigation across specific tasks are observed. We further explore the factors that may influence these phenomena such as artifacts in image generation (e.g. blurry faces) or pre-existing biases in the original datasets.
****
Bayes' Rays: Uncertainty Quantification for Neural Radiance Fields
Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20061-20070
Neural Radiance Fields (NeRFs) have shown promise in applications like view synt

hesis and depth estimation but learning from multiview images faces inherent unc
ertainties. Current methods to quantify them are either heuristic or computation
ally demanding. We introduce BayesRays a post-hoc framework to evaluate uncertai
nty in any pretrained NeRF without modifying the training process. Our method es
tablishes a volumetric uncertainty field using spatial perturbations and a Bayes
ian Laplace approximation. We derive our algorithm statistically and show its su
perior performance in key metrics and applications. Additional results available
 at: https://bayesrays.github.io/
****
NIVeL: Neural Implicit Vector Layers for Text-to-Vector Generation
Vikas Thamizharasan, Difan Liu, Matthew Fisher, Nanxuan Zhao, Evangelos Kalogera
kis, Michal Lukac; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2024, pp. 4589-4597
The success of denoising diffusion models in representing rich data distribution
s over 2D raster images has prompted research on extending them to other data re
presentations such as vector graphics. Unfortunately due to their variable struc
ture and scarcity of vector training data directly applying diffusion models on
this domain remains a challenging problem. Using workarounds like optimization v
ia Score Distillation Sampling (SDS) is also fraught with difficulty as vector r
epresentations are non-trivial to directly optimize and tend to result in implau
sible geometries such as redundant or self-intersecting shapes. NIVeL addresses
these challenges by reinterpreting the problem on an alternative intermediate do
main which preserves the desirable properties of vector graphics - mainly sparsi
ty of representation and resolution-independence. This alternative domain is bas
ed on neural implicit fields expressed in a set of decomposable editable layers.
 Based on our experiments NIVeL produces text-to-vector graphics results of sign
ificantly better quality than the state-of-the-art.
****
Driving-Video Dehazing with Non-Aligned Regularization for Safety Assistance
Junkai Fan, Jiangwei Weng, Kun Wang, Yijun Yang, Jianjun Qian, Jun Li, Jian Yang
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2024, pp. 26109-26119
Real driving-video dehazing poses a significant challenge due to the inherent di
fficulty in acquiring precisely aligned hazy/clear video pairs for effective mod
el training especially in dynamic driving scenarios with unpredictable weather c
onditions. In this paper we propose a pioneering approach that addresses this ch
allenge through a nonaligned regularization strategy. Our core concept involves
identifying clear frames that closely match hazy frames serving as references to
 supervise a video dehazing network. Our approach comprises two key components:
reference matching and video dehazing. Firstly we introduce a non-aligned refere
nce frame matching module leveraging an adaptive sliding window to match high-qu
ality reference frames from clear videos. Video dehazing incorporates flow-guide
d cosine attention sampler and deformable cosine attention fusion modules to enh
ance spatial multiframe alignment and fuse their improved information. To valida
te our approach we collect a GoProHazy dataset captured effortlessly with GoPro
cameras in diverse rural and urban road environments. Extensive experiments demo
nstrate the superiority of the proposed method over current state-of-the-art met
hods in the challenging task of real driving-video dehazing. Project page.
****
Is Vanilla MLP in Neural Radiance Field Enough for Few-shot View Synthesis?
Hanxin Zhu, Tianyu He, Xin Li, Bingchen Li, Zhibo Chen; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2028
8-20298
Neural Radiance Field (NeRF) has achieved superior performance for novel view sy
nthesis by modeling the scene with a Multi-Layer Perception (MLP) and a volume r
endering procedure however when fewer known views are given (i.e. few-shot view
synthesis) the model is prone to overfit the given views. To handle this issue p
revious efforts have been made towards leveraging learned priors or introducing
additional regularizations. In contrast in this paper we for the first time prov
ide an orthogonal method from the perspective of network structure. Given the ob

servation that trivially reducing the number of model parameters alleviates the overfitting issue but at the cost of missing details we propose the multi-input MLP (mi-MLP) that incorporates the inputs (i.e. location and viewing direction) of the vanilla MLP into each layer to prevent the overfitting issue without harming detailed synthesis. To further reduce the artifacts we propose to model colors and volume density separately and present two regularization terms. Extensive experiments on multiple datasets demonstrate that: 1) although the proposed mi-MLP is easy to implement it is surprisingly effective as it boosts the PSNR of the baseline from 14.73 to 24.23. 2) the overall framework achieves state-of-the-art results on a wide range of benchmarks. We will release the code upon publication.
****
CVT-xRF: Contrastive In-Voxel Transformer for 3D Consistent Radiance Fields from Sparse Inputs
Yingji Zhong, Lanqing Hong, Zhenguo Li, Dan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21466-21475
Neural Radiance Fields (NeRF) have shown impressive capabilities for photorealistic novel view synthesis when trained on dense inputs. However when trained on sparse inputs NeRF typically encounters issues of incorrect density or color predictions mainly due to insufficient coverage of the scene causing partial and sparse supervision thus leading to significant performance degradation. While existing works mainly consider ray-level consistency to construct 2D learning regularization based on rendered color depth or semantics on image planes in this paper we propose a novel approach that models 3D spatial field consistency to improve NeRF's performance with sparse inputs. Specifically we first adopt a voxel-based ray sampling strategy to ensure that the sampled rays intersect with a certain voxel in 3D space. We then randomly sample additional points within the voxel and apply a Transformer to infer the properties of other points on each ray which are then incorporated into the volume rendering. By backpropagating through the rendering loss we enhance the consistency among neighboring points. Additionally we propose to use a contrastive loss on the encoder output of the Transformer to further improve consistency within each voxel. Experiments demonstrate that our method yields significant improvement over different radiance fields in the sparse inputs setting and achieves comparable performance with current works. The project page for this paper is available at https://zhongyingji.github.io/CVT-xRF.
****
OAKINK2: A Dataset of Bimanual Hands-Object Manipulation in Complex Task Completion
Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 445-456
We present OAKINK2 a dataset of bimanual object manipulation tasks for complex daily activities. In pursuit of constructing the complex tasks into a structured representation OAKINK2 introduces three level of abstraction to organize the manipulation tasks: Affordance Primitive Task and Complex Task. OAKINK2 features on an object-centric perspective for decoding the complex tasks treating them as a sequence of object affordance fulfillment. The first level Affordance outlines the functionalities that objects in the scene can afford the second level Primitive Task describes the minimal interaction units that humans interact with the object to achieve its affordance and the third level Complex Task illustrates how Primitive Tasks are composed and interdependent. OAKINK2 dataset provides multi-view image streams and precise pose annotations for the human body hands and various interacting objects. This extensive collection supports applications such as interaction reconstruction and motion synthesis. Based on the 3-level abstraction of OAKINK2 we explore a task-oriented framework for Complex Task Completion (CTC). CTC aims to generate a sequence of bimanual manipulation to achieve task objectives. Within the CTC framework we employ Large Language Models (LLMs) to decompose the complex task objectives into sequences of Primitive Tasks and have developed a Motion Fulfillment Model that generates bimanual hand motion for ea

ch Primitive Task. OAKINK2 datasets and models are available at https://oakink.n
et/v2.
****
CogAgent: A Visual Language Model for GUI Agents
Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wa
ng, Zihan Wang, Yuxiao Dong, Ming Ding, Jie Tang; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14281-1429
0
People are spending an enormous amount of time on digital devices through graphi
cal user interfaces (GUIs) e.g. computer or smartphone screens. Large language m
odels (LLMs) such as ChatGPT can assist people in tasks like writing emails but
struggle to understand and interact with GUIs thus limiting their potential to i
ncrease automation levels. In this paper we introduce CogAgent an 18-billion-par
ameter visual language model (VLM) specializing in GUI understanding and navigat
ion. By utilizing both low-resolution and high-resolution image encoders CogAgen
t supports input at a resolution of 1120*1120 enabling it to recognize tiny page
 elements and text. As a generalist visual language model CogAgent achieves the
state of the art on five text-rich and four general VQA benchmarks including VQA
v2 OK-VQA Text-VQA ST-VQA ChartQA infoVQA DocVQA MM-Vet and POPE. CogAgent using
 only screenshots as input outperforms LLM-based methods that consume extracted
HTML text on both PC and Android GUI navigation tasks---Mind2Web and AITW advanc
ing the state of the art. The model and codes are available at https://github.co
m/THUDM/CogVLM.
****
Text-Guided 3D Face Synthesis - From Generation to Editing
Yunjie Wu, Yapeng Meng, Zhipeng Hu, Lincheng Li, Haoqian Wu, Kun Zhou, Weiwei Xu
, Xin Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2024, pp. 1260-1269
Text-guided 3D face synthesis has achieved remarkable results by leveraging text
-to-image (T2I) diffusion models. However most existing works focus solely on th
e direct generation ignoring the editing restricting them from synthesizing cust
omized 3D faces through iterative adjustments. In this paper we propose a unifie
d text-guided framework from face generation to editing. In the generation stage
 we propose a geometry-texture decoupled generation to mitigate the loss of geom
etric details caused by coupling. Besides decoupling enables us to utilize the g
enerated geometry as a condition for texture generation yielding highly geometry
-texture aligned results. We further employ a fine-tuned texture diffusion model
 to enhance texture quality in both RGB and YUV space. In the editing stage we f
irst employ a pre-trained diffusion model to update facial geometry or texture b
ased on the texts. To enable sequential editing we introduce a UV domain consist
ency preservation regularization preventing unintentional changes to irrelevant
facial attributes. Besides we propose a self-guided consistency weight strategy
to improve editing efficacy while preserving consistency. Through comprehensive
experiments we showcase our method's superiority in face synthesis.
****
AIDE: An Automatic Data Engine for Object Detection in Autonomous Driving
Mingfu Liang, Jong-Chyi Su, Samuel Schulter, Sparsh Garg, Shiyu Zhao, Ying Wu, M
anmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2024, pp. 14695-14706
Autonomous vehicle (AV) systems rely on robust perception models as a cornerston
e of safety assurance. However objects encountered on the road exhibit a long-ta
iled distribution with rare or unseen categories posing challenges to a deployed
 perception model. This necessitates an expensive process of continuously curati
ng and annotating data with significant human effort. We propose to leverage rec
ent advances in vision-language and large language models to design an Automatic
 Data Engine (AIDE) that automatically identifies issues efficiently curates dat
a improves the model through auto-labeling and verifies the model through genera
tion of diverse scenarios. This process operates iteratively allowing for contin
uous self-improvement of the model. We further establish a benchmark for open-wo
rld detection on AV datasets to comprehensively evaluate various learning paradi

gms demonstrating our method's superior performance at a reduced cost.
****
Multiplane Prior Guided Few-Shot Aerial Scene Rendering
Zihan Gao, Licheng Jiao, Lingling Li, Xu Liu, Fang Liu, Puhua Chen, Yuwei Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5009-5019
Neural Radiance Fields (NeRF) have been successfully applied in various aerial scenes yet they face challenges with sparse views due to limited supervision. The acquisition of dense aerial views is often prohibitive as unmanned aerial vehicles (UAVs) may encounter constraints in perspective range and energy constraints. In this work we introduce Multiplane Prior guided NeRF (MPNeRF) a novel approach tailored for few-shot aerial scene rendering--marking a pioneering effort in this domain. Our key insight is that the intrinsic geometric regularities specific to aerial imagery could be leveraged to enhance NeRF in sparse aerial scenes. By investigating NeRF's and Multiplane Image (MPI)'s behavior we propose to guide the training process of NeRF with a Multiplane Prior. The proposed Multiplane Prior draws upon MPI's benefits and incorporates advanced image comprehension through a SwinV2 Transformer pre-trained via SimMIM. Our extensive experiments demonstrate that MPNeRF outperforms existing state-of-the-art methods applied in non-aerial contexts by tripling the performance in SSIM and LPIPS even with three views available. We hope our work offers insights into the development of NeRF-based applications in aerial scenes with limited data.
****
MAS: Multi-view Ancestral Sampling for 3D Motion Generation Using 2D Diffusion
Roy Kapon, Guy Tevet, Daniel Cohen-Or, Amit H. Bermano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1965-1974
We introduce Multi-view Ancestral Sampling (MAS) a method for 3D motion generation using 2D diffusion models that were trained on motions obtained from in-the-wild videos. As such MAS opens opportunities to exciting and diverse fields of motion previously under-explored as 3D data is scarce and hard to collect. MAS works by simultaneously denoising multiple 2D motion sequences representing different views of the same 3D motion. It ensures consistency across all views at each diffusion step by combining the individual generations into a unified 3D sequence and projecting it back to the original views. We demonstrate MAS on 2D pose data acquired from videos depicting professional basketball maneuvers rhythmic gymnastic performances featuring a ball apparatus and horse races. In each of these domains 3D motion capture is arduous and yet MAS generates diverse and realistic 3D sequences. Unlike the Score Distillation approach which optimizes each sample by repeatedly applying small fixes our method uses a sampling process that was constructed for the diffusion framework. As we demonstrate MAS avoids common issues such as out-of-domain sampling and mode-collapse. https://guytevet.github.io/mas-page/
****
Smart Help: Strategic Opponent Modeling for Proactive and Adaptive Robot Assistance in Households
Zhihao Cao, Zidong Wang, Siwen Xie, Anji Liu, Lifeng Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18091-18101
Despite the significant demand for assistive technology among vulnerable groups (e.g. the elderly children and the disabled) in daily tasks research into advanced AI-driven assistive solutions that genuinely accommodate their diverse needs remains sparse. Traditional human-machine interaction tasks often require machines to simply help without nuanced consideration of human abilities and feelings such as their opportunity for practice and learning sense of self-improvement and self-esteem. Addressing this gap we define a pivotal and novel challenge Smart Help which aims to provide proactive yet adaptive support to human agents with diverse disabilities and dynamic goals in various tasks and environments. To establish this challenge we leverage AI2-THOR to build a new interactive 3D realistic household environment for the Smart Help task. We introduce an innovative opp

onent modeling module that provides a nuanced understanding of the main agent's capabilities and goals in order to optimize the assisting agent's helping policy. Rigorous experiments validate the efficacy of our model components and show the superiority of our holistic approach against established baselines. Our findings illustrate the potential of AI-imbued assistive robots in improving the well-being of vulnerable groups.
****
Bilateral Event Mining and Complementary for Event Stream Super-Resolution
Zhilin Huang, Quanmin Liang, Yijie Yu, Chujun Qin, Xiawu Zheng, Kai Huang, Zikun Zhou, Wenming Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 34-43
Event Stream Super-Resolution (ESR) aims to address the challenge of insufficient spatial resolution in event streams which holds great significance for the application of event cameras in complex scenarios. Previous works for ESR often process positive and negative events in a mixed paradigm. This paradigm limits their ability to effectively model the unique characteristics of each event and mutually refine each other by considering their correlations. In this paper we propose a bilateral event mining and complementary network (BMCNet) to fully leverage the potential of each event and capture the shared information to complement each other simultaneously. Specifically we resort to a two-stream network to accomplish comprehensive mining of each type of events individually. To facilitate the exchange of information between two streams we propose a bilateral information exchange (BIE) module. This module is layer-wisely embedded between two streams enabling the effective propagation of hierarchical global information while alleviating the impact of invalid information brought by inherent characteristics of events. The experimental results demonstrate that our approach outperforms the previous state-of-the-art methods in ESR achieving performance improvements of over 11% on both real and synthetic datasets. Moreover our method significantly enhances the performance of event-based downstream tasks such as object recognition and video reconstruction. Our code is available at https://github.com/Lqm26/BMCNet-ESR.
****
Online Task-Free Continual Generative and Discriminative Learning via Dynamic Cluster Memory
Fei Ye, Adrian G. Bors; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26202-26212
Online Task-Free Continual Learning (OTFCL) aims to learn novel concepts from streaming data without accessing task information. Most memory-based approaches used in OTFCL are not suitable for unsupervised learning because they require accessing supervised signals to implement their sample selection mechanisms. In this study we address this issue by proposing a novel memory management approach namely the Dynamic Cluster Memory (DCM) which builds new memory clusters to capture distribution shifts over time without accessing any supervised signals. DCM introduces a novel memory expansion mechanism based on the knowledge discrepancy criterion which evaluates the novelty of the incoming data as the signal for the memory expansion ensuring a compact memory capacity. We also propose a new sample selection approach that automatically stores incoming data samples with similar semantic information in the same memory cluster while also facilitating the knowledge diversity among memory clusters. Furthermore a novel memory pruning approach is proposed to automatically remove overlapping memory clusters through a graph relation evaluation ensuring a fixed memory capacity while maintaining the diversity among the samples stored in the memory. The proposed DCM is model-free plug-and-play and can be used in both supervised and unsupervised learning without modifications. Empirical results on OTFCL experiments show that the proposed DCM outperforms the state-of-the-art while requiring fewer data samples to be stored. The source code is available at https://github.com/dtuzi123/DCM.
****
Rapid Motor Adaptation for Robotic Manipulator Arms
Yichao Liang, Kevin Ellis, João Henriques; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16404-16413

Developing generalizable manipulation skills is a core challenge in embodied AI. This includes generalization across diverse task configurations encompassing variations in object shape density friction coefficient and external disturbances such as forces applied to the robot. Rapid Motor Adaptation (RMA) offers a promising solution to this challenge. It posits that essential hidden variables influencing an agent's task performance such as object mass and shape can be effectively inferred from the agent's action and proprioceptive history. Drawing inspiration from RMA in locomotion and in-hand rotation we use depth perception to develop agents tailored for rapid motor adaptation in a variety of manipulation tasks. We evaluated our agents on four challenging tasks from the Maniskill2 benchmark namely pick-and-place operations with hundreds of objects from the YCB and EGAD datasets peg insertion with precise position and orientation and operating a variety of faucets and handles with customized environment variations. Empirical results demonstrate that our agents surpass state-of-the-art methods like automatic domain randomization and vision-based policies obtaining better generalization performance and sample efficiency.
****
SANeRF-HQ: Segment Anything for NeRF in High Quality
Yichen Liu, Benran Hu, Chi-Keung Tang, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3216-3226
Recently the Segment Anything Model (SAM) has showcased remarkable capabilities of zero-shot segmentation while NeRF (Neural Radiance Fields) has gained popularity as a method for various 3D problems beyond novel view synthesis. Though there exist initial attempts to incorporate these two methods into 3D segmentation they face the challenge of accurately and consistently segmenting objects in complex scenarios. In this paper we introduce the Segment Anything for NeRF in High Quality (SANeRF-HQ) to achieve high-quality 3D segmentation of any target object in a given scene. SANeRF-HQ utilizes SAM for open-world object segmentation guided by user-supplied prompts while leveraging NeRF to aggregate information from different viewpoints. To overcome the aforementioned challenges we employ density field and RGB similarity to enhance the accuracy of segmentation boundary during the aggregation. Emphasizing on segmentation accuracy we evaluate our method on multiple NeRF datasets where high-quality ground-truths are available or manually annotated. SANeRF-HQ shows a significant quality improvement over state-of-the-art methods in NeRF object segmentation provides higher flexibility for object localization and enables more consistent object segmentation across multiple views.
****
DSGG: Dense Relation Transformer for an End-to-end Scene Graph Generation
Zeeshan Hayder, Xuming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28317-28326
Scene graph generation aims to capture detailed spatial and semantic relationships between objects in an image which is challenging due to incomplete labeling long-tailed relationship categories and relational semantic overlap. Existing Transformer-based methods either employ distinct queries for objects and predicates or utilize holistic queries for relation triplets and hence often suffer from limited capacity in learning low-frequency relationships. In this paper we present a new Transformer-based method called DSGG that views scene graph detection as a direct graph prediction problem based on a unique set of graph-aware queries. In particular each graph-aware query encodes a compact representation of both the node and all of its relations in the graph acquired through the utilization of a relaxed sub-graph matching during the training process. Moreover to address the problem of relational semantic overlap we utilize a strategy for relation distillation aiming to efficiently learn multiple instances of semantic relationships. Extensive experiments on the VG and the PSG datasets show that our model achieves state-of-the-art results showing a significant improvement of 3.5% and 6.7% in mR@50 and mR@100 for the scene-graph generation task and achieves an even more substantial improvement of 8.5% and 10.3% in mR@50 and mR@100 for the panoptic scene graph generation task. Code is available at https://github.com/zeeshan

hayder/DSGG.
****
Transcending the Limit of Local Window: Advanced Super-Resolution Transformer with Adaptive Token Dictionary
Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, Shuhang Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2856-2865
Single Image Super-Resolution is a classic computer vision problem that involves estimating high-resolution (HR) images from low-resolution (LR) ones. Although deep neural networks (DNNs) especially Transformers for super-resolution have seen significant advancements in recent years challenges still remain particularly in limited receptive field caused by window-based self-attention. To address these issues we introduce a group of auxiliary Adaptive Token Dictionary to SR Transformer and establish an ATD-SR method. The introduced token dictionary could learn prior information from training data and adapt the learned prior to specific testing image through an adaptive refinement step. The refinement strategy could not only provide global information to all input tokens but also group image tokens into categories. Based on category partitions we further propose a category-based self-attention mechanism designed to leverage distant but similar tokens for enhancing input features. The experimental results show that our method achieves the best performance on various single image super-resolution benchmarks.
****
Object Dynamics Modeling with Hierarchical Point Cloud-based Representations
Chanho Kim, Li Fuxin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20977-20986
Modeling object dynamics with a neural network is an important problem with numerous applications. Most recent work has been based on graph neural networks. However physics happens in 3D space where geometric information potentially plays an important role in modeling physical phenomena. In this work we propose a novel U-net architecture based on continuous point convolution which naturally embeds information from 3D coordinates and allows for multi-scale feature representations with established downsampling and upsampling procedures. Bottleneck layers in the downsampled point clouds lead to better long-range interaction modeling. Besides the flexibility of point convolutions allows our approach to generalize to sparsely sampled points from mesh vertices and dynamically generate features on important interaction points on mesh faces. Experimental results demonstrate that our approach significantly improves the state-of-the-art especially in scenarios that require accurate gravity or collision reasoning.
****
WWW: A Unified Framework for Explaining What Where and Why of Neural Networks by Interpretation of Neuron Concepts
Yong Hyun Ahn, Hyeon Bae Kim, Seong Tae Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10968-10977
Recent advancements in neural networks have showcased their remarkable capabilities across various domains. Despite these successes the "black box" problem still remains. To address this we propose a novel framework WWW that offers the 'what' 'where' and 'why' of the neural network decisions in human-understandable terms. Specifically WWW utilizes an adaptive selection for concept discovery employing adaptive cosine similarity and thresholding techniques to effectively explain 'what'. To address the 'where' and 'why' we proposed a novel combination of neuron activation maps (NAMs) with Shapley values generating localized concept maps and heatmaps for individual inputs. Furthermore WWW introduces a method for predicting uncertainty leveraging heatmap similarities to estimate the prediction's reliability. Experimental evaluations of WWW demonstrate superior performance in both quantitative and qualitative metrics outperforming existing methods in interpretability. WWW provides a unified solution for explaining 'what' 'where' and 'why' introducing a method for localized explanations from global interpretations and offering a plug-and-play solution adaptable to various architectures. The code is available at: https://github.com/ailab-kyunghee/WWW
****

SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interp retation for Earth Observation Imagery
Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong , Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, Yansheng Li; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2024, pp. 27672-27683
Prior studies on Remote Sensing Foundation Model (RSFM) reveal immense potential towards a generic model for Earth Observation. Nevertheless these works primari ly focus on a single modality without temporal and geo-context modeling hamperin g their capabilities for diverse tasks. In this study we present SkySense a gene ric billion-scale model pre-trained on a curated multi-modal Remote Sensing Imag ery (RSI) dataset with 21.5 million temporal sequences. SkySense incorporates a factorized multi-modal spatiotemporal encoder taking temporal sequences of optic al and Synthetic Aperture Radar (SAR) data as input. This encoder is pre-trained by our proposed Multi-Granularity Contrastive Learning to learn representations across different modal and spatial granularities. To further enhance the RSI re presentations by the geo-context clue we introduce Geo-Context Prototype Learnin g to learn region-aware prototypes upon RSI's multi-modal spatiotemporal feature s. To our best knowledge SkySense is the largest Multi-Modal RSFM to date whose modules can be flexibly combined or used individually to accommodate various tas ks. It demonstrates remarkable generalization capabilities on a thorough evaluat ion encompassing 16 datasets over 7 tasks from single- to multi-modal static to temporal and classification to localization. SkySense surpasses 18 recent RSFMs in all test scenarios. Specifically it outperforms the latest models such as GFM SatLas and Scale-MAE by a large margin i.e. 2.76% 3.67% and 3.61% on average re spectively. We will release the pre-trained weights to facilitate future researc h and Earth Observation applications.
****
CaKDP: Category-aware Knowledge Distillation and Pruning Framework for Lightweig ht 3D Object Detection
Haonan Zhang, Longjun Liu, Yuqi Huang, Zhao Yang, Xinyu Lei, Bihan Wen; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024, pp. 15331-15341
Knowledge distillation (KD) possesses immense potential to accelerate the deep n eural networks (DNNs) for LiDAR-based 3D detection. However in most of prevailin g approaches the suboptimal teacher models and insufficient student architecture investigations limit the performance gains. To address these issues we propose a simple yet effective Category-aware Knowledge Distillation and Pruning (CaKDP) framework for compressing 3D detectors. Firstly CaKDP transfers the knowledge o f two-stage detector to one-stage student one mitigating the impact of inadequat e teacher models. To bridge the gap between the heterogeneous detectors we inves tigate their differences and then introduce the student-motivated category-aware KD to align the category prediction between distillation pairs. Secondly we pro pose a category-aware pruning scheme to obtain the customizable architecture of compact student model. The method calculates the category prediction gap before and after removing each filter to evaluate the importance of filters and retains the important filters. Finally to further improve the student performance a mod ified IOU-aware refinement module with negligible computations is leveraged to r emove the redundant false positive predictions. Experiments demonstrate that CaK DP achieves the compact detector with high performance. For example on WOD CaKDP accelerates CenterPoint by half while boosting L2 mAPH by 1.61%. The code is av ailable at https://github.com/zhnxjtu/CaKDP.
****
Mixed-Precision Quantization for Federated Learning on Resource-Constrained Hete rogeneous Devices
Huancheng Chen, Haris Vikalo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6138-6148
While federated learning (FL) systems often utilize quantization to battle commu nication and computational bottlenecks they have heretofore been limited to depl oying fixed-precision quantization schemes. Meanwhile the concept of mixed-preci

sion quantization (MPQ) where different layers of a deep learning model are assi gned varying bit-width remains unexplored in the FL settings. We present a novel FL algorithm FedMPQ which introduces mixed-precision quantization to resource-h eterogeneous FL systems. Specifically local models quantized so as to satisfy bi t-width constraint are trained by optimizing an objective function that includes a regularization term which promotes reduction of precision in some of the laye rs without significant performance degradation. The server collects local model updates de-quantizes them into full-precision models and then aggregates them in to a global model. To initialize the next round of local training the server rel ies on the information learned in the previous training round to customize bit-w idth assignments of the models delivered to different clients. In extensive benc hmarking experiments on several model architectures and different datasets in bo th iid and non-iid settings FedMPQ outperformed the baseline FL schemes that uti lize fixed-precision quantization while incurring only a minor computational ove rhead on the participating devices.
****
CFAT: Unleashing Triangular Windows for Image Super-resolution
Abhisek Ray, Gaurav Kumar, Maheshkumar H. Kolekar; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26120-261 29
Transformer-based models have revolutionized the field of image super-resolution (SR) by harnessing their inherent ability to capture complex contextual feature s. The overlapping rectangular shifted window technique used in transformer arch itecture nowadays is a common practice in super-resolution models to improve the quality and robustness of image upscaling. However it suffers from distortion a t the boundaries and has limited unique shifting modes. To overcome these weakne sses we propose a non-overlapping triangular window technique that synchronously works with the rectangular one to mitigate boundary-level distortion and allows the model to access more unique sifting modes. In this paper we propose a Compo site Fusion Attention Transformer (CFAT) that incorporates triangular-rectangula r window-based local attention with a channel-based global attention technique i n image super-resolution. As a result CFAT enables attention mechanisms to be ac tivated on more image pixels and captures long-range multi-scale features to imp rove SR performance. The extensive experimental results and ablation study demon strate the effectiveness of CFAT in the SR domain. Our proposed model shows a si gnificant 0.7 dB performance improvement over other state-of-the-art SR architec tures.
****
ICP-Flow: LiDAR Scene Flow Estimation with ICP
Yancong Lin, Holger Caesar; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2024, pp. 15501-15511
Scene flow characterizes the 3D motion between two LiDAR scans captured by an au tonomous vehicle at nearby timesteps. Prevalent methods consider scene flow as p oint-wise unconstrained flow vectors that can be learned by either large-scale t raining beforehand or time-consuming optimization at inference. However these me thods do not take into account that objects in autonomous driving often move rig idly. We incorporate this rigid-motion assumption into our design where the goal is to associate objects over scans and then estimate the locally rigid transfor mations. We propose ICP-Flow a learning-free flow estimator. The core of our des ign is the conventional Iterative Closest Point (ICP) algorithm which aligns the objects over time and outputs the corresponding rigid transformations. Cruciall y to aid ICP we propose a histogram-based initialization that discovers the most likely translation thus providing a good starting point for ICP. The complete s cene flow is then recovered from the rigid transformations. We outperform state- of-the-art baselines including supervised models on the Waymo dataset and perfor m competitively on Argoverse-v2 and nuScenes. Further we train a feedforward neu ral network supervised by the pseudo labels from our model and achieve top perfo rmance among all models capable of real-time inference. We validate the advantag e of our model on scene flow estimation with longer temporal gaps up to 0.4 seco nds where other models fail to deliver meaningful results.

****
MADTP: Multimodal Alignment-Guided Dynamic Token Pruning for Accelerating Vision-Language Transformer
Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, Tao Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15710-15719
Vision-Language Transformers (VLTs) have shown great success recently but are meanwhile accompanied by heavy computation costs where a major reason can be attributed to the large number of visual and language tokens. Existing token pruning research for compressing VLTs mainly follows a single-modality-based scheme yet ignores the critical role of aligning different modalities for guiding the token pruning process causing the important tokens for one modality to be falsely pruned in another modality branch. Meanwhile existing VLT pruning works also lack the flexibility to dynamically compress each layer based on different input samples. To this end we propose a novel framework named Multimodal Alignment-Guided Dynamic Token Pruning (MADTP) for accelerating various VLTs. Specifically we first introduce a well-designed Multi-modality Alignment Guidance (MAG) module that can align features of the same semantic concept from different modalities to ensure the pruned tokens are less important for all modalities. We further design a novel Dynamic Token Pruning (DTP) module which can adaptively adjust the token compression ratio in each layer based on different input instances. Extensive experiments on various benchmarks demonstrate that MADTP significantly reduces the computational complexity of kinds of multimodal models while preserving competitive performance. Notably when applied to the BLIP model in the NLVR2 dataset MADTP can reduce the GFLOPs by 80% with less than 4% performance degradation.
****
G-NeRF: Geometry-enhanced Novel View Synthesis from Single-View Images
Zixiong Huang, Qi Chen, Libo Sun, Yifan Yang, Naizhou Wang, Qi Wu, Mingkui Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10117-10126
Novel view synthesis aims to generate new view images of a given view image collection. Recent attempts address this problem relying on 3D geometry priors (e.g. shapes sizes and positions) learned from multi-view images. However such methods encounter the following limitations: 1) they require a set of multi-view images as training data for a specific scene (e.g. face car or chair) which is often unavailable in many real-world scenarios; 2) they fail to extract the geometry priors from single-view images due to the lack of multi-view supervision. In this paper we propose a Geometry-enhanced NeRF (G-NeRF) which seeks to enhance the geometry priors by a geometry-guided multi-view synthesis approach followed by a depth-aware training. In the synthesis process inspired that existing 3D GAN models can unconditionally synthesize high-fidelity multi-view images we seek to adopt off-the-shelf 3D GAN models such as EG3D as a free source to provide geometry priors through synthesizing multi-view data. Simultaneously to further improve the geometry quality of the synthetic data we introduce a truncation method to effectively sample latent codes within 3D GAN models. To tackle the absence of multi-view supervision for single-view images we design the depth-aware training approach incorporating a depth-aware discriminator to guide geometry priors through depth maps. Experiments demonstrate the effectiveness of our method in terms of both qualitative and quantitative results.
****
Neural Fields as Distributions: Signal Processing Beyond Euclidean Space
Daniel Rebain, Soroosh Yazdani, Kwang Moo Yi, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4274-4283
Neural fields have emerged as a powerful and broadly applicable method for representing signals. However in contrast to classical discrete digital signal processing the portfolio of tools to process such representations is still severely limited and restricted to Euclidean domains. In this paper we address this problem by showing how a probabilistic re-interpretation of neural fields can enable their training and inference processes to become "filter-aware". The formulation w

e propose not only merges training and filtering in an efficient way but also ge
neralizes beyond the familiar Euclidean coordinate spaces to the more general se
t of smooth manifolds and convolutions induced by the actions of Lie groups. We
demonstrate how this framework can enable novel integrations of signal processin
g techniques for neural field applications on both Euclidean domains such as ima
ges and audio as well as non-Euclidean domains such as rotations and rays. A not
eworthy benefit of our method is its applicability. Our method can be summarized
 as primarily a modification of the loss function and in most cases does not req
uire changes to the network architecture or the inference process.
****
Rolling Shutter Correction with Intermediate Distortion Flow Estimation
Mingdeng Cao, Sidi Yang, Yujiu Yang, Yinqiang Zheng; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25338-2
5347
This paper proposes to correct the rolling shutter (RS) distorted images by esti
mating the distortion flow from the global shutter (GS) to RS directly. Existing
 methods usually perform correction using the undistortion flow from the RS to G
S. They initially predict the flow from consecutive RS frames subsequently resca
ling it as the displacement fields from the RS frame to the underlying GS image
using time-dependent scaling factors. Following this RS-aware forward warping is
 employed to convert the RS image into its GS counterpart. Nevertheless this str
ategy is prone to two shortcomings. First the undistortion flow estimation is re
ndered inaccurate by merely linear scaling the flow due to the complex non-linea
r motion nature. Second RS-aware forward warping often results in unavoidable ar
tifacts. To address these limitations we introduce a new framework that directly
 estimates the distortion flow and rectifies the RS image with the backward warp
ing operation. More specifically we first propose a global correlation-based flo
w attention mechanism to estimate the initial distortion flow and GS feature joi
ntly which are then refined by the following coarse-to-fine decoder layers. Addi
tionally a multi-distortion flow prediction strategy is integrated to mitigate t
he issue of inaccurate flow estimation further. Experimental results validate th
e effectiveness of the proposed method which outperforms state-of-the-art approa
ches on various benchmarks while maintaining high efficiency. The project is ava
ilable at https://github.com/ljzycmd/DFRSC.
****
Style Blind Domain Generalized Semantic Segmentation via Covariance Alignment an
d Semantic Consistence Contrastive Learning
Woo-Jin Ahn, Geun-Yeong Yang, Hyun-Duck Choi, Myo-Taeg Lim; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
3616-3626
Deep learning models for semantic segmentation often experience performance degr
adation when deployed to unseen target domains unidentified during the training
phase. This is mainly due to variations in image texture (i.e. style) from diffe
rent data sources. To tackle this challenge existing domain generalized semantic
 segmentation (DGSS) methods attempt to remove style variations from the feature
. However these approaches struggle with the entanglement of style and content w
hich may lead to the unintentional removal of crucial content information causin
g performance degradation. This study addresses this limitation by proposing Bli
ndNet a novel DGSS approach that blinds the style without external modules or da
tasets. The main idea behind our proposed approach is to alleviate the effect of
 style in the encoder whilst facilitating robust segmentation in the decoder. To
 achieve this BlindNet comprises two key components: covariance alignment and se
mantic consistency contrastive learning. Specifically the covariance alignment t
rains the encoder to uniformly recognize various styles and preserve the content
 information of the feature rather than removing the style-sensitive factor. Mea
nwhile semantic consistency contrastive learning enables the decoder to construc
t discriminative class embedding space and disentangles features that are vulner
able to misclassification. Through extensive experiments our approach outperform
s existing DGSS methods exhibiting robustness and superior performance for seman
tic segmentation on unseen target domains.

****
Attack To Defend: Exploiting Adversarial Attacks for Detecting Poisoned Models
Samar Fares, Karthik Nandakumar; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2024, pp. 24726-24735
Poisoning (trojan/backdoor) attacks enable an adversary to train and deploy a co
rrupted machine learning (ML) model which typically works well and achieves good
 accuracy on clean input samples but behaves maliciously on poisoned samples con
taining specific trigger patterns. Using such poisoned ML models as the foundati
on to build real-world systems can compromise application safety. Hence there is
 a critical need for algorithms that detect whether a given target model has bee
n poisoned. This work proposes a novel approach for detecting poisoned models ca
lled Attack To Defend (A2D) which is based on the observation that poisoned mode
ls are more sensitive to adversarial perturbations compared to benign models. We
 propose a metric called sensitivity to adversarial perturbations (SAP) to measu
re the sensitivity of a ML model to adversarial attacks at a specific perturbati
on bound. We then generate strong adversarial attacks against an unrelated refer
ence model and estimate the SAP value of the target model by transferring the ge
nerated attacks. The target model is deemed to be a trojan if its SAP value exce
eds a decision threshold. The A2D framework requires only black-box access to th
e target model and a small clean set while being computationally efficient. The
A2D approach has been evaluated on four standard image datasets and its effectiv
eness under various types of poisoning attacks has been demonstrated
****
X-3D: Explicit 3D Structure Modeling for Point Cloud Recognition
Shuofeng Sun, Yongming Rao, Jiwen Lu, Haibin Yan; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5074-5083
Numerous prior studies predominantly emphasize constructing relation vectors for
 individual neighborhood points and generating dynamic kernels for each vector a
nd embedding these into high-dimensional spaces to capture implicit local struct
ures. However we contend that such implicit high-dimensional structure modeling
approch inadequately represents the local geometric structure of point clouds du
e to the absence of explicit structural information. Hence we introduce X-3D an
explicit 3D structure modeling approach. X-3D functions by capturing the explici
t local structural information within the input 3D space and employing it to pro
duce dynamic kernels with shared weights for all neighborhood points within the
current local region. This modeling approach introduces effective geometric prio
r and significantly diminishes the disparity between the local structure of the
embedding space and the original input point cloud thereby improving the extract
ion of local features. Experiments show that our method can be used on a variety
 of methods and achieves state-of-the-art performance on segmentation classifica
tion detection tasks with lower extra computational cost. Such as 90.7% on ScanO
bjectNN for classification 79.2% on S3DIS 6 fold and 74.3% on S3DIS Area 5 for s
egmentation 76.3% on ScanNetV2 for segmentation and 64.5% mAP_ 25 46.9% mAP_ 50
  on SUN RGB-D and 69.0% mAP_ 25 51.1% mAP_ 50 on ScanNetV2 . Our code is avai
lable at \href https://github.com/sunshuofeng/X-3D https://github.com/sunshuofe
ng/X-3D .
****
SpiderMatch: 3D Shape Matching with Global Optimality and Geometric Consistency
Paul Roetzer, Florian Bernard; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2024, pp. 14543-14553
Finding shortest paths on product spaces is a popular approach to tackle numerou
s variants of matching problems including the dynamic time warping method for ma
tching signals the matching of curves or the matching of a curve to a 3D shape.
While these approaches admit the computation of globally optimal solutions in po
lynomial time their natural generalisation to 3D shape matching is widely known
to be intractable. In this work we address this issue by proposing a novel path-
based formalism for 3D shape matching. More specifically we consider an alternat
ive shape discretisation in which one of the 3D shapes (the source shape) is rep
resented as a SpiderCurve i.e. a long self-intersecting curve that traces the 3D
 shape surface. We then tackle the 3D shape matching problem as finding a shorte

st path in the product graph of the SpiderCurve and the target 3D shape. Our app roach introduces a set of novel constraints that ensure a globally geometrically consistent matching. Overall our formalism leads to an integer linear programming problem for which we experimentally show that it can efficiently be solved to global optimality. We demonstrate that our approach is competitive with recent state-of-the-art shape matching methods while in addition guaranteeing geometric consistency.
****
Troika: Multi-Path Cross-Modal Traction for Compositional Zero-Shot Learning
Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, Donglin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24005-24014
Recent compositional zero-shot learning (CZSL) methods adapt pre-trained vision-language models (VLMs) by constructing trainable prompts only for composed state-object pairs. Relying on learning the joint representation of seen compositions these methods ignore the explicit modeling of the state and object thus limiting the exploitation of pre-trained knowledge and generalization to unseen compositions. With a particular focus on the universality of the solution in this work we propose a novel paradigm for CZSL models that establishes three identification branches (i.e. Multi-Path) to jointly model the state object and composition. The presented Troika is an outstanding implementation that aligns the branch-specific prompt representations with decomposed visual features. To calibrate the bias between semantically similar multi-modal representations we further devise a Cross-Modal Traction module into Troika that shifts the prompt representation towards the current visual content. We conduct extensive experiments on three popular benchmarks where our method significantly outperforms existing methods in both closed-world and open-world settings. The code will be available at https://github.com/bighuang624/Troika.
****
One More Step: A Versatile Plug-and-Play Module for Rectifying Diffusion Schedule Flaws and Enhancing Low-Frequency Controls
Minghui Hu, Jianbin Zheng, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, Tat-Jen Cham; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7331-7340
It is well known that many open-released foundational diffusion models have difficulty in generating images that substantially depart from average brightness despite such images being present in the training data. This is due to an inconsistency: while denoising starts from pure Gaussian noise during inference the training noise schedule retains residual data even in the final timestep distribution due to difficulties in numerical conditioning in mainstream formulation leading to unintended bias during inference. To mitigate this issue certain eps-prediction models are combined with an ad-hoc offset-noise methodology. In parallel some contemporary models have adopted zero-terminal SNR noise schedules together with v-prediction which necessitate major alterations to pre-trained models. However such changes risk destabilizing a large multitude of community-driven applications anchored on these pre-trained models. In light of this our investigation revisits the fundamental causes leading to our proposal of an innovative and principled remedy called One More Step (OMS). By integrating a compact network and incorporating an additional simple yet effective step during inference OMS elevates image fidelity and harmonizes the dichotomy between training and inference while preserving original model parameters. Once trained various pre-trained diffusion models with the same latent domain can share the same OMS module.
****
Enhancing Multimodal Cooperation via Sample-level Modality Valuation
Yake Wei, Ruoxuan Feng, Zihe Wang, Di Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27338-27347
One primary topic of multimodal learning is to jointly incorporate heterogeneous information from different modalities. However most models often suffer from unsatisfactory multimodal cooperation which cannot jointly utilize all modalities well. Some methods are proposed to identify and enhance the worse learnt modalit

y but they are often hard to provide the fine-grained observation of multimodal cooperation at sample-level with theoretical support. Hence it is essential to reasonably observe and improve the fine-grained cooperation between modalities especially when facing realistic scenarios where the modality discrepancy could vary across different samples. To this end we introduce a sample-level modality valuation metric to evaluate the contribution of each modality for each sample. Via modality valuation we observe that modality discrepancy indeed could be different at sample-level beyond the global contribution discrepancy at dataset-level. We further analyze this issue and improve cooperation between modalities at sample-level by enhancing the discriminative ability of low-contributing modalities in a targeted manner. Overall our methods reasonably observe the fine-grained uni-modal contribution and achieve considerable improvement. The source code and dataset are available at https://github.com/GeWu-Lab/Valuate-and-Enhance-Multimodal-Cooperation.
****

Evidential Active Recognition: Intelligent and Prudent Open-World Embodied Perception

Lei Fan, Mingfu Liang, Yunxuan Li, Gang Hua, Ying Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16351-16361

Active recognition enables robots to intelligently explore novel observations thereby acquiring more information while circumventing undesired viewing conditions. Recent approaches favor learning policies from simulated or collected data wherein appropriate actions are more frequently selected when the recognition is accurate. However most recognition modules are developed under the closed-world assumption which makes them ill-equipped to handle unexpected inputs such as the absence of the target object in the current observation. To address this issue we propose treating active recognition as a sequential evidence-gathering process providing by-step uncertainty quantification and reliable prediction under the evidence combination theory. Additionally the reward function developed in this paper effectively characterizes the merit of actions when operating in open-world environments. To evaluate the performance we collect a dataset from an indoor simulator encompassing various recognition challenges such as distance occlusion levels and visibility. Through a series of experiments on recognition and robustness analysis we demonstrate the necessity of introducing uncertainties to active recognition and the superior performance of the proposed method.
****

SatSynth: Augmenting Image-Mask Pairs through Diffusion Models for Aerial Semantic Segmentation

Aysim Toker, Marvin Eisenberger, Daniel Cremers, Laura Leal-Taixé; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27695-27705

In recent years semantic segmentation has become a pivotal tool in processing and interpreting satellite imagery. Yet a prevalent limitation of supervised learning techniques remains the need for extensive manual annotations by experts. In this work we explore the potential of generative image diffusion to address the scarcity of annotated data in earth observation tasks. The main idea is to learn the joint data manifold of images and labels leveraging recent advancements in denoising diffusion probabilistic models. To the best of our knowledge we are the first to generate both images and corresponding masks for satellite segmentation. We find that the obtained pairs not only display high quality in fine-scale features but also ensure a wide sampling diversity. Both aspects are crucial for earth observation data where semantic classes can vary severely in scale and occurrence frequency. We employ the novel data instances for downstream segmentation as a form of data augmentation. In our experiments we provide comparisons to prior works based on discriminative diffusion models or GANs. We demonstrate that integrating generated samples yields significant quantitative improvements for satellite semantic segmentation -- both compared to baselines and when training only on the original data.
****

XScale-NVS: Cross-Scale Novel View Synthesis with Hash Featurized Manifold
Guangyu Wang, Jinzhi Zhang, Fan Wang, Ruqi Huang, Lu Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21029-21039
We propose XScale-NVS for high-fidelity cross-scale novel view synthesis of real-world large-scale scenes. Existing representations based on explicit surface suffer from discretization resolution or UV distortion while implicit volumetric representations lack scalability for large scenes due to the dispersed weight distribution and surface ambiguity. In light of the above challenges we introduce hash featurized manifold a novel hash-based featurization coupled with a deferred neural rendering framework. This approach fully unlocks the expressivity of the representation by explicitly concentrating the hash entries on the 2D manifold thus effectively representing highly detailed contents independent of the discretization resolution. We also introduce a novel dataset namely GigaNVS to benchmark cross-scale high-resolution novel view synthesis of real-world large-scale scenes. Our method significantly outperforms competing baselines on various real-world scenes yielding an average LPIPS that is ?40% lower than prior state-of-the-art on the challenging GigaNVS benchmark. Please see our project page at: xscalenvs.github.io.
****
Ink Dot-Oriented Differentiable Optimization for Neural Image Halftoning
Hao Jiang, Bingfeng Zhou, Yadong Mu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27528-27537
Halftoning is a time-honored printing technique that simulates continuous tones using ink dots (halftone dots). The resurgence of deep learning has catalyzed the emergence of innovative technologies in the printing industry fostering the advancement of data-driven halftoning methods. Nevertheless current deep learning-based approaches produce halftones through image-to-image black box transformations lacking direct control over the movement of individual halftone dots. In this paper we propose an innovative halftoning method termed "neural dot-controllable halftoning". This method allows dot-level image dithering by providing direct control over the motion of each ink dot. We conceptualize halftoning as the process of sprinkling dots on a canvas. Initially a specific quantity of dots are randomly dispersed on the canvas and subsequently adjusted based on the surrounding grayscale and gradient. To establish differentiable transformations between discrete ink dot positions and halftone matrices we devise a lightweight dot encoding network to spread dense gradients to sparse dots. Dot control offers several advantages to our approach including the capability to regulate the quantity of halftone dots and enhance specific areas with artifacts in the generated halftones by adjusting the placement of the dots. Our proposed method exhibits superior performance than previous approaches in extensive quantitative and qualitative experiments.
****
The Unreasonable Effectiveness of Pre-Trained Features for Camera Pose Refinement
Gabriele Trivigno, Carlo Masone, Barbara Caputo, Torsten Sattler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12786-12798
Pose refinement is an interesting and practically relevant research direction. Pose refinement can be used to (1) obtain a more accurate pose estimate from an initial prior (e.g. from retrieval) (2) as pre-processing i.e. to provide a better starting point to a more expensive pose estimator (3) as post-processing of a more accurate localizer. Existing approaches focus on learning features / scene representations for the pose refinement task. This involves training an implicit scene representation or learning features while optimizing a camera pose-based loss. A natural question is whether training specific features / representations is truly necessary or whether similar results can be already achieved with more generic features. In this work we present a simple approach that combines pre-trained features with a particle filter and a renderable representation of the scene. Despite its simplicity it achieves state-of-the-art results demonstrating t

hat one can easily build a pose refiner without the need for specific training. The code will be released upon acceptance.
****
Scalable 3D Registration via Truncated Entry-wise Absolute Residuals
Tianyu Huang, Liangzu Peng, Rene Vidal, Yun-Hui Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27477-27487
Given an input set of 3D point pairs the goal of outlier-robust 3D registration is to compute some rotation and translation that align as many point pairs as possible. This is an important problem in computer vision for which many highly accurate approaches have been recently proposed. Despite their impressive performance these approaches lack scalability often overflowing the 16GB of memory of a standard laptop to handle roughly 30000 point pairs. In this paper we propose a 3D registration approach that can process more than ten million (10^7) point pairs with over 99% random outliers. Moreover our method is efficient entails low memory costs and maintains high accuracy at the same time. We call our method TEAR as it involves minimizing an outlier-robust loss that computes Truncated Entry-wise Absolute Residuals. To minimize this loss we decompose the original 6-dimensional problem into two subproblems of dimensions 3 and 2 respectively solved in succession to global optimality via a customized branch-and-bound method. While branch-and-bound is often slow and unscalable this does not apply to TEAR as we propose novel bounding functions that are tight and computationally efficient. Experiments on various datasets are conducted to validate the scalability and efficiency of our method.
****
ExtraNeRF: Visibility-Aware View Extrapolation of Neural Radiance Fields with Diffusion Models
Meng-Li Shih, Wei-Chiu Ma, Lorenzo Boyice, Aleksander Holynski, Forrester Cole, Brian Curless, Janne Kontkanen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20385-20395
We propose ExtraNeRF a novel method for extrapolating the range of views handled by a Neural Radiance Field (NeRF). Our main idea is to leverage NeRFs to model scene-specific fine-grained details while capitalizing on diffusion models to extrapolate beyond our observed data. A key ingredient is to track visibility to determine what portions of the scene have not been observed and focus on reconstructing those regions consistently with diffusion models. Our primary contributions include a visibility-aware diffusion-based inpainting module that is fine-tuned on the input imagery yielding an initial NeRF with moderate quality (often blurry) inpainted regions followed by a second diffusion model trained on the input imagery to consistently enhance notably sharpen the inpainted imagery from the first pass. We demonstrate high-quality results extrapolating beyond a small number of (typically six or fewer) input views effectively outpainting the NeRF as well as inpainting newly disoccluded regions inside the original viewing volume. We compare with related work both quantitatively and qualitatively and show significant gains over prior art.
****
Equivariant Plug-and-Play Image Reconstruction
Matthieu Terris, Thomas Moreau, Nelly Pustelnik, Julian Tachella; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25255-25264
Plug-and-play algorithms constitute a popular framework for solving inverse imaging problems that rely on the implicit definition of an image prior via a denoiser. These algorithms can leverage powerful pre-trained denoisers to solve a wide range of imaging tasks circumventing the necessity to train models on a per-task basis. Unfortunately plug-and-play methods often show unstable behaviors hampering their promise of versatility and leading to suboptimal quality of reconstructed images. In this work we show that enforcing equivariance to certain groups of transformations (rotations reflections and/or translations) on the denoiser strongly improves the stability of the algorithm as well as its reconstruction quality. We provide a theoretical analysis that illustrates the role of equivarian

ce on better performance and stability. We present a simple algorithm that enforces equivariance on any existing denoiser by simply applying a random transformation to the input of the denoiser and the inverse transformation to the output at each iteration of the algorithm. Experiments on multiple imaging modalities and denoising networks show that the equivariant plug-and-play algorithm improves both the reconstruction performance and the stability compared to their non-equivariant counterparts.
****
CLIP as RNN: Segment Countless Visual Concepts without Training Endeavor
Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, Siyang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13171-13182
Existing open-vocabulary image segmentation methods require a fine-tuning step on mask labels and/or image-text datasets. Mask labels are labor-intensive which limits the number of categories in segmentation datasets. Consequently the vocabulary capacity of pre-trained VLMs is severely reduced after fine-tuning. However without fine-tuning VLMs trained under weak image-text supervision tend to make suboptimal mask predictions. To alleviate these issues we introduce a novel recurrent framework that progressively filters out irrelevant texts and enhances mask quality without training efforts. The recurrent unit is a two-stage segmenter built upon a frozen VLM. Thus our model retains the VLM's broad vocabulary space and equips it with segmentation ability. Experiments show that our method outperforms not only the training-free counterparts but also those fine-tuned with millions of data samples and sets the new state-of-the-art records for both zero-shot semantic and referring segmentation. Concretely we improve the current record by 28.8 16.0 and 6.9 mIoU on Pascal VOC COCO Object and Pascal Context.
****
LP++: A Surprisingly Strong Linear Probe for Few-Shot CLIP
Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, Ismail Ben Ayed; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23773-23782
In a recent strongly emergent literature on few-shot CLIP adaptation Linear Probe (LP) has been often reported as a weak baseline. This has motivated intensive research building convoluted prompt learning or feature adaptation strategies. In this work we propose and examine from convex-optimization perspectives a generalization of the standard LP baseline in which the linear classifier weights are learnable functions of the text embedding with class-wise multipliers blending image and text knowledge. As our objective function depends on two types of variables i.e. the class visual prototypes and the learnable blending parameters we propose a computationally efficient block coordinate Majorize-Minimize (MM) descent algorithm. In our full-batch MM optimizer which we coin LP++ step sizes are implicit unlike standard gradient descent practices where learning rates are intensively searched over validation sets. By examining the mathematical properties of our loss (e.g. Lipschitz gradient continuity) we build majorizing functions yielding data-driven learning rates and derive approximations of the loss's minima which provide data-informed initialization of the variables. Our image-language objective function along with these non-trivial optimization insights and ingredients yields surprisingly highly competitive few-shot CLIP performances. Furthermore LP++ operates in black-box relaxes intensive validation searches for the optimization hyper-parameters and runs orders-of-magnitudes faster than state-of-the-art few-shot CLIP adaptation methods. Our code is available at: https://github.com/FereshteShakeri/FewShot-CLIP-Strong-Baseline.git.
****
Active Generalized Category Discovery
Shijie Ma, Fei Zhu, Zhun Zhong, Xu-Yao Zhang, Cheng-Lin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16890-16900
Generalized Category Discovery (GCD) is a pragmatic and challenging open-world task which endeavors to cluster unlabeled samples from both novel and old classes leveraging some labeled data of old classes. Given that knowledge learned from

old classes is not fully transferable to new classes and that novel categories are fully unlabeled GCD inherently faces intractable problems including imbalanced classification performance and inconsistent confidence between old and new classes especially in the low-labeling regime. Hence some annotations of new classes are deemed necessary. However labeling new classes is extremely costly. To address this issue we take the spirit of active learning and propose a new setting called Active Generalized Category Discovery (AGCD). The goal is to improve the performance of GCD by actively selecting a limited amount of valuable samples for labeling from the oracle. To solve this problem we devise an adaptive sampling strategy which jointly considers novelty informativeness and diversity to adaptively select novel samples with proper uncertainty. However owing to the varied orderings of label indices caused by the clustering of novel classes the queried labels are not directly applicable to subsequent training. To overcome this issue we further propose a stable label mapping algorithm that transforms ground truth labels to the label space of the classifier thereby ensuring consistent training across different active selection stages. Our method achieves state-of-the-art performance on both generic and fine-grained datasets. Our code is available at https://github.com/mashijie1028/ActiveGCD

****

HIVE: Harnessing Human Feedback for Instructional Visual Editing

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, Ran Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9026-9036

Incorporating human feedback has been shown to be crucial to align text generated by large language models to human preferences. We hypothesize that state-of-the-art instructional image editing models where outputs are generated based on an input image and an editing instruction could similarly benefit from human feedback as their outputs may not adhere to the correct instructions and preferences of users. In this paper we present a novel framework to harness human feedback for instructional visual editing (HIVE). Specifically we collect human feedback on the edited images and learn a reward function to capture the underlying user preferences. We then introduce scalable diffusion model fine-tuning methods that can incorporate human preferences based on the estimated reward. Besides to mitigate the bias brought by the limitation of data we contribute a new 1.1M training dataset a 3.6K reward dataset for rewards learning and a 1K evaluation dataset to boost the performance of instructional image editing. We conduct extensive empirical experiments quantitatively and qualitatively showing that HIVE is favored over previous state-of-the-art instructional image editing approaches by a large margin.

****

StrokeFaceNeRF: Stroke-based Facial Appearance Editing in Neural Radiance Field

Xiao-Juan Li, Dingxi Zhang, Shu-Yu Chen, Feng-Lin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7538-7547

Current 3D-aware facial NeRF generation approaches control the facial appearance by text lighting conditions or reference images limiting precise manipulation of local facial regions and interactivity. Color stroke a user-friendly and effective tool to depict appearance is challenging to edit 3D faces because of the lack of texture coarse geometry representation and detailed editing operations. To solve the above problems we introduce StrokeFaceNeRF a novel stroke-based method for editing facial NeRF appearance. In order to infer the missing texture and 3D geometry information 2D edited stroke maps are firstly encoded into the EG3D's latent space followed by a transformer-based editing module to achieve effective appearance changes while preserving the original geometry in editing regions. Notably we design a novel geometry loss function to ensure surface density remains consistent during training. To further enhance the local manipulation accuracy we propose a stereo fusion approach which lifts the 2D mask (inferred from strokes or drawn by users) into 3D mask volume allowing explicit blending of the original and edited faces. Extensive experiments validate that the proposed metho

d outperforms existing 2D and 3D methods in both editing reality and geometry re tention.
****
FlowVQTalker: High-Quality Emotional Talking Face Generation through Normalizing Flow and Quantization
Shuai Tan, Bin Ji, Ye Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26317-26327
Generating emotional talking faces is a practical yet challenging endeavor. To create a lifelike avatar we draw upon two critical insights from a human perspective: 1) The connection between audio and the non-deterministic facial dynamics encompassing expressions blinks poses should exhibit synchronous and one-to-many mapping. 2) Vibrant expressions are often accompanied by emotion-aware high-definition (HD) textures and finely detailed teeth. However both aspects are frequently overlooked by existing methods. To this end this paper proposes using normalizing Flow and Vector-Quantization modeling to produce emotional talking faces that satisfy both insights concurrently (FlowVQTalker). Specifically we develop a flowbased coefficient generator that encodes the dynamics of facial emotion into a multi-emotion-class latent space represented as a mixture distribution. The generation process commences with random sampling from the modeled distribution guided by the accompanying audio enabling both lip-synchronization and the uncertain nonverbal facial cues generation. Furthermore our designed vector-quantization image generator treats the creation of expressive facial images as a code query task utilizing a learned codebook to provide rich high-quality textures that enhance the emotional perception of the results. Extensive experiments are conducted to showcase the effectiveness of our approach.
****
Learning from Observer Gaze: Zero-Shot Attention Prediction Oriented by Human-Object Interaction Recognition
Yuchen Zhou, Linkai Liu, Chao Gou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28390-28400
Most existing attention prediction research focuses on salient instances like humans and objects. However the more complex interaction-oriented attention arising from the comprehension of interactions between instances by human observers remains largely unexplored. This is equally crucial for advancing human-machine interaction and human-centered artificial intelligence. To bridge this gap we first collect a novel gaze fixation dataset named IG comprising 530000 fixation points across 740 diverse interaction categories capturing visual attention during human observers' cognitive processes of interactions. Subsequently we introduce the zero-shot interaction-oriented attention prediction task (ZeroIA) which challenges models to predict visual cues for interactions not encountered during training. Thirdly we present the Interactive Attention model (IA) designed to emulate human observers' cognitive processes to tackle the ZeroIA problem. Extensive experiments demonstrate that the proposed IA outperforms other state-of-the-art approaches in both ZeroIA and fully supervised settings. Lastly we endeavor to apply interaction-oriented attention to the interaction recognition task itself. Further experimental results demonstrate the promising potential to enhance the performance and interpretability of existing state-of-the-art HOI models by incorporating real human attention data from IG and attention labels generated by IA.
****
ProxyCap: Real-time Monocular Full-body Capture in World Space via Human-Centric Proxy-to-Motion Learning
Yuxiang Zhang, Hongwen Zhang, Liangxiao Hu, Jiajun Zhang, Hongwei Yi, Shengping Zhang, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1954-1964
Learning-based approaches to monocular motion capture have recently shown promising results by learning to regress in a data-driven manner. However due to the challenges in data collection and network designs it remains challenging to achieve real-time full-body capture while being accurate in world space. In this work we introduce ProxyCap a human-centric proxy-to-motion learning scheme to learn world-space motions from a proxy dataset of 2D skeleton sequences and 3D rotatio

nal motions. Such proxy data enables us to build a learning-based network with a
ccurate world-space supervision while also mitigating the generalization issues.
 For more accurate and physically plausible predictions in world space our netwo
rk is designed to learn human motions from a human-centric perspective which ena
bles the understanding of the same motion captured with different camera traject
ories. Moreover a contact-aware neural motion descent module is proposed to impr
ove foot-ground contact and motion misalignment with the proxy observations. Wit
h the proposed learning-based solution we demonstrate the first real-time monocu
lar full-body capture system with plausible foot-ground contact in world space e
ven using hand-held cameras.
****
OpenBias: Open-set Bias Detection in Text-to-Image Generative Models
Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqia
n Xu, Zhangyang Wang, Humphrey Shi, Nicu Sebe; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12225-12235
Text-to-image generative models are becoming increasingly popular and accessible
 to the general public. As these models see large-scale deployments it is necess
ary to deeply investigate their safety and fairness to not disseminate and perpe
tuate any kind of biases. However existing works focus on detecting closed sets
of biases defined a priori limiting the studies to well-known concepts. In this
paper we tackle the challenge of open-set bias detection in text-to-image genera
tive models presenting OpenBias a new pipeline that identifies and quantifies th
e severity of biases agnostically without access to any precompiled set. OpenBia
s has three stages. In the first phase we leverage a Large Language Model (LLM)
to propose biases given a set of captions. Secondly the target generative model
produces images using the same set of captions. Lastly a Vision Question Answeri
ng model recognizes the presence and extent of the previously proposed biases. W
e study the behavior of Stable Diffusion 1.5 2 and XL emphasizing new biases nev
er investigated before. Via quantitative experiments we demonstrate that OpenBia
s agrees with current closed-set bias detection methods and human judgement.
****
On the Robustness of Language Guidance for Low-Level Vision Tasks: Findings from
 Depth Estimation
Agneet Chatterjee, Tejas Gokhale, Chitta Baral, Yezhou Yang; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 2794-2803
Recent advances in monocular depth estimation have been made by incorporating na
tural language as additional guidance. Although yielding impressive results the
impact of the language prior particularly in terms of generalization and robustn
ess remains unexplored. In this paper we address this gap by quantifying the imp
act of this prior and introduce methods to benchmark its effectiveness across va
rious settings. We generate "low-level" sentences that convey object-centric thr
ee-dimensional spatial relationships incorporate them as additional language pri
ors and evaluate their downstream impact on depth estimation. Our key finding is
 that current language-guided depth estimators perform optimally only with scene
-level descriptions and counter-intuitively fare worse with low level descriptio
ns. Despite leveraging additional data these methods are not robust to directed
adversarial attacks and decline in performance with an increase in distribution
shift. Finally to provide a foundation for future research we identify points of
 failures and offer insights to better understand these shortcomings. With an in
creasing number of methods using language for depth estimation our findings high
light the opportunities and pitfalls that require careful consideration for effe
ctive deployment in real-world settings.
****
UFOGen: You Forward Once Large Scale Text-to-Image Generation via Diffusion GANs
Yanwu Xu, Yang Zhao, Zhisheng Xiao, Tingbo Hou; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8196-8206
Text-to-image diffusion models have demonstrated remarkable capabilities in tran
sforming text prompts into coherent images yet the computational cost of the mul
ti-step inference remains a persistent challenge. To address this issue we prese

nt UFOGen a novel generative model designed for ultra-fast one-step text-to-imag
e generation. In contrast to conventional approaches that focus on improving sam
plers or employing distillation techniques for diffusion models UFOGen adopts a
hybrid methodology integrating diffusion models with a GAN objective. Leveraging
 a newly introduced diffusion-GAN objective and initialization with pre-trained
diffusion models UFOGen excels in efficiently generating high-quality images con
ditioned on textual descriptions in a single step. Beyond traditional text-to-im
age generation UFOGen showcases versatility in applications. Notably UFOGen stan
ds among the pioneering models enabling one-step text-to-image generation and di
verse downstream tasks presenting a significant advancement in the landscape of
efficient generative models.
****
3DiffTection: 3D Object Detection with Geometry-Aware Diffusion Features
Chenfeng Xu, Huan Ling, Sanja Fidler, Or Litany; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10617-10627
3DiffTection introduces a novel method for 3D object detection from single image
s utilizing a 3D-aware diffusion model for feature extraction. Addressing the re
source-intensive nature of annotating large-scale 3D image data our approach lev
erages pretrained diffusion models traditionally used for 2D tasks and adapts th
em for 3D detection through geometric and semantic tuning. Geometrically we enha
nce the model to perform view synthesis from single images incorporating an epip
olar warp operator. This process utilizes easily accessible posed image data eli
minating the need for manual annotation. Semantically the model is further refin
ed on target detection data. Both stages utilize ControlNet ensuring the preserv
ation of original feature capabilities. Through our methodology we obtain 3D-awa
re features that excel in identifying cross-view point correspondences. In 3D de
tection 3DiffTection substantially surpasses previous benchmarks e.g. Cube-RCNN
by 9.43% in AP3D on the Omni3D-ARkitscene dataset. Furthermore 3DiffTection demo
nstrates robust label efficiency and generalizes well to cross-domain data nearl
y matching fully-supervised models in zero-shot scenarios.
****
Lift3D: Zero-Shot Lifting of Any 2D Vision Model to 3D
Mukund Varma T, Peihao Wang, Zhiwen Fan, Zhangyang Wang, Hao Su, Ravi Ramamoorth
i; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2024, pp. 21367-21377
In recent years there has been an explosion of 2D vision models for numerous tas
ks such as semantic segmentation style transfer or scene editing enabled by larg
e-scale 2D image datasets. At the same time there has been renewed interest in 3
D scene representations such as neural radiance fields from multi-view images. H
owever the availability of 3D or multiview data is still substantially limited c
ompared to 2D image datasets making extending 2D vision models to 3D data highly
 desirable but also very challenging. Indeed extending a single 2D vision operat
or like scene editing to 3D typically requires a highly creative method speciali
zed to that task and often requires per-scene optimization. In this paper we ask
 the question of whether any 2D vision model can be lifted to make 3D consistent
 predictions. We answer this question in the affirmative; our new Lift3D method
trains to predict unseen views on feature spaces generated by a few visual model
s (i.e. DINO and CLIP) but then generalizes to novel vision operators and tasks
such as style transfer super-resolution open vocabulary segmentation and image c
olorization; for some of these tasks there is no comparable previous 3D method.
In many cases we even outperform state-of-the-art methods specialized for the ta
sk in question. Moreover Lift3D is a zero-shot method in the sense that it requi
res no task-specific training nor scene-specific optimization.
****
LowRankOcc: Tensor Decomposition and Low-Rank Recovery for Vision-based 3D Seman
tic Occupancy Prediction
Linqing Zhao, Xiuwei Xu, Ziwei Wang, Yunpeng Zhang, Borui Zhang, Wenzhao Zheng,
Dalong Du, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2024, pp. 9806-9815
In this paper we present a tensor decomposition and low-rank recovery approach (

LowRankOcc) for vision-based 3D semantic occupancy prediction. Conventional meth ods model outdoor scenes with fine-grained 3D grids but the sparsity of non-empt y voxels introduces considerable spatial redundancy leading to potential overfit ting risks. In contrast our approach leverages the intrinsic low-rank property o f 3D occupancy data factorizing voxel representations into low-rank components t o efficiently mitigate spatial redundancy without sacrificing performance. Speci fically we present the Vertical-Horizontal (VH) decomposition block factorizes 3 D tensors into vertical vectors and horizontal matrices. With our "decomposition -encoding-recovery" framework we encode 3D contexts with only 1/2D convolutions and poolings and subsequently recover the encoded compact yet informative contex t features back to voxel representations. Experimental results demonstrate that LowRankOcc achieves state-of-the-art performances in semantic scene completion o n the SemanticKITTI dataset and 3D occupancy prediction on the nuScenes dataset.
****

Multiway Point Cloud Mosaicking with Diffusion and Global Optimization
Shengze Jin, Iro Armeni, Marc Pollefeys, Daniel Barath; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2083 8-20849
We introduce a novel framework for multiway point cloud mosaicking (named Wednes day) designed to co-align sets of partially overlapping point clouds -- typicall y obtained from 3D scanners or moving RGB-D cameras -- into a unified coordinate system. At the core of our approach is ODIN a learned pairwise registration alg orithm that iteratively identifies overlaps and refines attention scores employi ng a diffusion-based process for denoising pairwise correlation matrices to enha nce matching accuracy. Further steps include constructing a pose graph from all point clouds performing rotation averaging a novel robust algorithm for re-estim ating translations optimally in terms of consensus maximization and translation optimization. Finally the point cloud rotations and positions are optimized join tly by a diffusion-based approach. Tested on four diverse large-scale datasets o ur method achieves state-of-the-art pairwise and multiway registration results b y a large margin on all benchmarks. Our code and models are available at https:/ /github.com/jinsz/Multiway-Point-Cloud-Mosaicking-with-Diffusion-and-Global-Opti mization.
****

Novel View Synthesis with View-Dependent Effects from a Single Image
Juan Luis Gonzalez Bello, Munchurl Kim; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10413-10423
In this paper we address single image-based novel view synthesis (NVS) by firstl y integrating view-dependent effects (VDE) into the process. Our approach levera ges camera motion priors to model VDE treating negative disparity as the represe ntation of these effects in the scene. By identifying that specularities align w ith camera motion we infuse VDEs into input images by aggregating pixel colors a long the negative depth region of epipolar lines. Additionally we introduce a re laxed volumetric rendering approximation enhancing efficiency by computing densi ties in a single pass for NVS from single images. Notably our method learns sing le-image NVS from image sequences alone making it a fully self-supervised learni ng approach that requires no depth or camera pose annotations. We present extens ive experimental results and show that our proposed method can learn NVS with VD Es outperforming the SOTA single-view NVS methods on the RealEstate10k and Manne quinChallenge datasets. Visit our project site https://kaist-viclab.github.io/mo novde-site.
****

Point2RBox: Combine Knowledge from Synthetic Visual Patterns for End-to-end Orie nted Object Detection with Single Point Supervision
Yi Yu, Xue Yang, Qingyun Li, Feipeng Da, Jifeng Dai, Yu Qiao, Junchi Yan; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2024, pp. 16783-16793
With the rapidly increasing demand for oriented object detection (OOD) recent re search involving weakly-supervised detectors for learning rotated box (RBox) fro m the horizontal box (HBox) has attracted more and more attention. In this paper

we explore a more challenging yet label-efficient setting namely single point-s
upervised OOD and present our approach called Point2RBox. Specifically we propos
e to leverage two principles: 1) Synthetic pattern knowledge combination: By sam
pling around each labeled point on the image we spread the object feature to syn
thetic visual patterns with known boxes to provide the knowledge for box regress
ion. 2) Transform self-supervision: With a transformed input image (e.g. scaled/
rotated) the output RBoxes are trained to follow the same transformation so that
 the network can perceive the relative size/rotation between objects. The detect
or is further enhanced by a few devised techniques to cope with peripheral issue
s e.g. the anchor/layer assignment as the size of the object is not available in
 our point supervision setting. To our best knowledge Point2RBox is the first en
d-to-end solution for point-supervised OOD. In particular our method uses a ligh
tweight paradigm yet it achieves a competitive performance among point-supervise
d alternatives 41.05%/27.62%/80.01% on DOTA/DIOR/HRSC datasets.
****
PBWR: Parametric-Building-Wireframe Reconstruction from Aerial LiDAR Point Cloud
s
Shangfeng Huang, Ruisheng Wang, Bo Guo, Hongxin Yang; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27778-
27787
In this paper we present an end-to-end 3D-building-wireframe reconstruction meth
od to regress edges directly from aerial light-detection-and-ranging (LiDAR) poi
nt clouds. Our method named parametric-building-wireframe reconstruction (PBWR)
takes aerial LiDAR point clouds and initial edge entities as input and fully use
s the self-attention mechanism of transformers to regress edge parameters withou
t any intermediate steps such as corner prediction. We propose an edge non-maxim
um suppression (E-NMS) module based on edge similarity to remove redundant edges
. Additionally a dedicated edge loss function is utilized to guide the PBWR in r
egressing edges parameters when the simple use of the edge distance loss is not
suitable. In our experiments our proposed method demonstrated state-of-the-art r
esults on the Building3D dataset achieving an improvement of approximately 36% i
n Entry-level dataset edge accuracy and around a 42% improvement in the Tallinn
dataset.
****
Spectrum AUC Difference (SAUCD): Human-aligned 3D Shape Evaluation
Tianyu Luan, Zhong Li, Lele Chen, Xuan Gong, Lichang Chen, Yi Xu, Junsong Yuan;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2024, pp. 20155-20164
Existing 3D mesh shape evaluation metrics mainly focus on the overall shape but
are usually less sensitive to local details. This makes them inconsistent with h
uman evaluation as human perception cares about both overall and detailed shape.
 In this paper we propose an analytic metric named Spectrum Area Under the Curve
 Difference (SAUCD) that demonstrates better consistency with human evaluation.
To compare the difference between two shapes we first transform the 3D mesh to t
he spectrum domain using the discrete Laplace-Beltrami operator and Fourier tran
sform. Then we calculate the Area Under the Curve (AUC) difference between the t
wo spectrums so that each frequency band that captures either the overall or det
ailed shape is equitably considered. Taking human sensitivity across frequency b
ands into account we further extend our metric by learning suitable weights for
each frequency band which better aligns with human perception. To measure the pe
rformance of SAUCD we build a 3D mesh evaluation dataset called Shape Grading al
ong with manual annotations from more than 800 subjects. By measuring the correl
ation between our metric and human evaluation we demonstrate that SAUCD is well
aligned with human evaluation and outperforms previous 3D mesh metrics.
****
HRVDA: High-Resolution Visual Document Assistant
Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jian
g, Xing Sun, Linli Xu; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2024, pp. 15534-15545
Leveraging vast training data multimodal large language models (MLLMs) have demo

nstrated formidable general visual comprehension capabilities and achieved remarkable performance across various tasks. However their performance in visual document understanding still leaves much room for improvement. This discrepancy is primarily attributed to the fact that visual document understanding is a fine-grained prediction task. In natural scenes MLLMs typically use low-resolution images leading to a substantial loss of visual information. Furthermore general-purpose MLLMs do not excel in handling document-oriented instructions. In this paper we propose a High-Resolution Visual Document Assistant (HRVDA) which bridges the gap between MLLMs and visual document understanding. This model employs a content filtering mechanism and an instruction filtering module to separately filter out the content-agnostic visual tokens and instruction-agnostic visual tokens thereby achieving efficient model training and inference for high-resolution images. In addition we construct a document-oriented visual instruction tuning dataset and apply a multi-stage training strategy to enhance the model's document modeling capabilities. Extensive experiments demonstrate that our model achieves state-of-the-art performance across multiple document understanding datasets while maintaining training efficiency and inference speed comparable to low-resolution models.
****
Learning for Transductive Threshold Calibration in Open-World Recognition
Qin Zhang, Dongsheng An, Tianjun Xiao, Tong He, Qingming Tang, Ying Nian Wu, Joseph Tighe, Yifan Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17097-17106
In deep metric learning for visual recognition the calibration of distance thresholds is crucial for achieving desired model performance in the true positive rates (TPR) or true negative rates (TNR). However calibrating this thresh- old presents challenges in open-world scenarios where the test classes can be entirely disjoint from those encountered during training. We define the problem of finding distance thresholds for a trained embedding model to achieve target performance metrics over unseen open-world test classes as open-world threshold calibration. Existing posthoc threshold calibration methods reliant on inductive inference and requiring a calibration dataset with a similar distance distribution as the test data often prove ineffective in open- world scenarios. To address this we introduce OpenGCN a Graph Neural Network-based transductive threshold calibration method with enhanced adaptability and robustness. OpenGCN learns to predict pairwise connectivity for the unlabeled test instances embedded in a graph to determine its TPR and TNR at various distance thresholds allowing for transductive inference of the distance thresholds which also incorporates test-time information. Extensive experiments across open-world visual recognition benchmarks validate OpenGCN's superiority over existing posthoc calibration methods for open-world threshold calibration.
****
Weakly-Supervised Emotion Transition Learning for Diverse 3D Co-speech Gesture Generation
Xingqun Qi, Jiahao Pan, Peng Li, Ruibin Yuan, Xiaowei Chi, Mengfei Li, Wenhan Luo, Wei Xue, Shanghang Zhang, Qifeng Liu, Yike Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10424-10434
Generating vivid and emotional 3D co-speech gestures is crucial for virtual avatar animation in human-machine interaction applications. While the existing methods enable generating the gestures to follow a single emotion label they overlook that long gesture sequence modeling with emotion transition is more practical in real scenes. In addition the lack of large-scale available datasets with emotional transition speech and corresponding 3D human gestures also limits the addressing of this task. To fulfill this goal we first incorporate the ChatGPT-4 and an audio inpainting approach to construct the high-fidelity emotion transition human speeches. Considering obtaining the realistic 3D pose annotations corresponding to the dynamically inpainted emotion transition audio is extremely difficult we propose a novel weakly supervised training strategy to encourage authority gesture transitions. Specifically to enhance the coordination of transition gest

ures w.r.t. different emotional ones we model the temporal association representation between two different emotional gesture sequences as style guidance and infuse it into the transition generation. We further devise an emotion mixture mechanism that provides weak supervision based on a learnable mixed emotion label for transition gestures. Last we present a keyframe sampler to supply effective initial posture cues in long sequences enabling us to generate diverse gestures. Extensive experiments demonstrate that our method outperforms the state-of-the-art models constructed by adapting single emotion-conditioned counterparts on our newly defined emotion transition task and datasets. Our code and dataset will be released on the project page: https://xingqunqi-lab.github.io/Emo-Transition-Gesture/
****
Multi-Session SLAM with Differentiable Wide-Baseline Pose Optimization
Lahav Lipson, Jia Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19626-19635
We introduce a new system for Multi-Session SLAM which tracks camera motion across multiple disjoint videos under a single global reference. Our approach couples the prediction of optical flow with solver layers to estimate camera pose. The backbone is trained end-to-end using a novel differentiable solver for wide-baseline two-view pose. The full system can connect disjoint sequences perform visual odometry and global optimization. Compared to existing approaches our design is accurate and robust to catastrophic failures.
****
A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation
Qucheng Peng, Ce Zheng, Chen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2240-2249
3D human pose data collected in controlled laboratory settings present challenges for pose estimators that generalize across diverse scenarios. To address this domain generalization is employed. Current methodologies in domain generalization for 3D human pose estimation typically utilize adversarial training to generate synthetic poses for training. Nonetheless these approaches exhibit several limitations. First the lack of prior information about the target domain complicates the application of suitable augmentation through a single pose augmentor affecting generalization on target domains. Moreover adversarial training's discriminator tends to enforce similarity between source and synthesized poses impeding the exploration of out-of-source distributions. Furthermore the pose estimator's optimization is not exposed to domain shifts limiting its overall generalization ability. To address these limitations we propose a novel framework featuring two pose augmentors: the weak and the strong augmentors. Our framework employs differential strategies for generation and discrimination processes facilitating the preservation of knowledge related to source poses and the exploration of out-of-source distributions without prior information about target poses. Besides we leverage meta-optimization to simulate domain shifts in the optimization process of the pose estimator thereby improving its generalization ability. Our proposed approach significantly outperforms existing methods as demonstrated through comprehensive experiments on various benchmark datasets.
****
Improving Out-of-Distribution Generalization in Graphs via Hierarchical Semantic Environments
Yinhua Piao, Sangseon Lee, Yijingxiu Lu, Sun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27631-27640
Out-of-distribution (OOD) generalization in the graph domain is challenging due to complex distribution shifts and a lack of environmental contexts. Recent methods attempt to enhance graph OOD generalization by generating flat environments. However such flat environments come with inherent limitations to capture more complex data distributions. Considering the DrugOOD dataset which contains diverse training environments (e.g. scaffold size etc.) flat contexts cannot sufficiently address its high heterogeneity. Thus a new challenge is posed to generate more semantically enriched environments to enhance graph invariant learning for ha

ndling distribution shifts. In this paper we propose a novel approach to generate hierarchical semantic environments for each graph. Firstly given an input graph we explicitly extract variant subgraphs from the input graph to generate proxy predictions on local environments. Then stochastic attention mechanisms are employed to re-extract the subgraphs for regenerating global environments in a hierarchical manner. In addition we introduce a new learning objective that guides our model to learn the diversity of environments within the same hierarchy while maintaining consistency across different hierarchies. This approach enables our model to consider the relationships between environments and facilitates robust graph invariant learning. Extensive experiments on real-world graph data have demonstrated the effectiveness of our framework. Particularly in the challenging dataset DrugOOD our method achieves up to 1.29% and 2.83% improvement over the best baselines on IC50 and EC50 prediction tasks respectively.
****
CN-RMA: Combined Network with Ray Marching Aggregation for 3D Indoor Object Detection from Multi-view Images
Guanlin Shen, Jingwei Huang, Zhihua Hu, Bin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21326-21335
This paper introduces CN-RMA a novel approach for 3D indoor object detection from multi-view images. We observe the key challenge as the ambiguity of image and 3D correspondence without explicit geometry to provide occlusion information. To address this issue CN-RMA leverages the synergy of 3D reconstruction networks and 3D object detection networks where the reconstruction network provides a rough Truncated Signed Distance Function (TSDF) and guides image features to vote to 3D space correctly in an end-to-end manner. Specifically we associate weights to sampled points of each ray through ray marching representing the contribution of a pixel in an image to corresponding 3D locations. Such weights are determined by the predicted signed distances so that image features vote only to regions near the reconstructed surface. Our method achieves state-of-the-art performance in 3D object detection from multi-view images as measured by mAP@0.25 and mAP@0.5 on the ScanNet and ARKitScenes datasets. The code and models are released at https://github.com/SerCharles/CN-RMA.
****
ACT-Diffusion: Efficient Adversarial Consistency Training for One-step Diffusion Models
Fei Kong, Jinhao Duan, Lichao Sun, Hao Cheng, Renjing Xu, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, Kaidi Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8890-8899
Though diffusion models excel in image generation their step-by-step denoising leads to slow generation speeds. Consistency training addresses this issue with single-step sampling but often produces lower-quality generations and requires high training costs. In this paper we show that optimizing consistency training loss minimizes the Wasserstein distance between target and generated distributions. As timestep increases the upper bound accumulates previous consistency training losses. Therefore larger batch sizes are needed to reduce both current and accumulated losses. We propose Adversarial Consistency Training (ACT) which directly minimizes the Jensen-Shannon (JS) divergence between distributions at each timestep using a discriminator. Theoretically ACT enhances generation quality and convergence. By incorporating a discriminator into the consistency training framework our method achieves improved FID scores on CIFAR10 and ImageNet 64x64 and LSUN Cat 256x256 datasets retains zero-shot image inpainting capabilities and uses less than 1/6 of the original batch size and fewer than 1/2 of the model parameters and training steps compared to the baseline method this leads to a substantial reduction in resource consumption. Our code is available: https://github.com/kong13661/ACT
****
Spectral Meets Spatial: Harmonising 3D Shape Matching and Interpolation
Dongliang Cao, Marvin Eisenberger, Nafie El Amrani, Daniel Cremers, Florian Bernard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.

nition (CVPR), 2024, pp. 3658-3668
Although 3D shape matching and interpolation are highly interrelated they are often studied separately and applied sequentially to relate different 3D shapes thus resulting in sub-optimal performance. In this work we present a unified framework to predict both point-wise correspondences and shape interpolation between 3D shapes. To this end we combine the deep functional map framework with classical surface deformation models to map shapes in both spectral and spatial domains. On the one hand by incorporating spatial maps our method obtains more accurate and smooth point-wise correspondences compared to previous functional map methods for shape matching. On the other hand by introducing spectral maps our method gets rid of commonly used but computationally expensive geodesic distance constraints that are only valid for near-isometric shape deformations. Furthermore we propose a novel test-time adaptation scheme to capture both pose-dominant and shape-dominant deformations. Using different challenging datasets we demonstrate that our method outperforms previous state-of-the-art methods for both shape matching and interpolation even compared to supervised approaches.
****
Emu Edit: Precise Image Editing via Recognition and Generation Tasks
Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, Yaniv Taigman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8871-8879
Instruction-based image editing holds immense potential for a variety of applications as it enables users to perform any editing operation using a natural language instruction. However current models in this domain often struggle with accurately executing user instructions. We present Emu Edit a multi-task image editing model which sets state-of-the-art results in instruction-based image editing. To develop Emu Edit we train it to multi-task across an unprecedented range of tasks such as region-based editing free-form editing and Computer Vision tasks all of which are formulated as generative tasks. Additionally to enhance Emu Edit's multi-task learning abilities we provide it with learned task embeddings which guide the generation process towards the correct edit type. Both these elements are essential for Emu Edit's outstanding performance. Furthermore we show that Emu Edit can generalize to new tasks such as image inpainting super-resolution and compositions of editing tasks with just a few labeled examples. This capability offers a significant advantage in scenarios where high-quality samples are scarce. Lastly to facilitate a more rigorous and informed assessment of instructable image editing models we release a new challenging and versatile benchmark that includes seven different image editing tasks.
****
Face2Diffusion for Fast and Editable Face Personalization
Kaede Shiohara, Toshihiko Yamasaki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6850-6859
Face personalization aims to insert specific faces taken from images into pretrained text-to-image diffusion models. However it is still challenging for previous methods to preserve both the identity similarity and editability due to overfitting to training samples. In this paper we propose Face2Diffusion (F2D) for high-editability face personalization. The core idea behind F2D is that removing identity-irrelevant information from the training pipeline prevents the overfitting problem and improves editability of encoded faces. F2D consists of the following three novel components: 1) Multi-scale identity encoder provides well-disentangled identity features while keeping the benefits of multi-scale information which improves the diversity of camera poses. 2) Expression guidance disentangles face expressions from identities and improves the controllability of face expressions. 3) Class-guided denoising regularization encourages models to learn how faces should be denoised which boosts the text-alignment of backgrounds. Extensive experiments on the FaceForensics++ dataset and diverse prompts demonstrate our method greatly improves the trade-off between the identity- and text-fidelity compared to previous state-of-the-art methods. Code is available at https://github.com/mapooon/Face2Diffusion.
****

Causal-CoG: A Causal-Effect Look at Context Generation for Boosting Multi-modal Language Models
Shitian Zhao, Zhuowan Li, Yadong Lu, Alan Yuille, Yan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13342-13351
While Multi-modal Language Models (MLMs) demon strate impressive multimodal ability they still struggle on providing factual and precise responses for tasks like vi sual question answering (VQA). In this paper we address this challenge from the perspective of contextual informa tion. We propose Causal Context Generation Causal-CoG which is a prompting strategy that engages contextual infor mation to enhance precise VQA during inference. Specifi cally we prompt MLMs to generate contexts i.e text de scription of an image and engage the generated contexts for question answering. Moreover we investigate the ad vantage of contexts on VQA from a causality perspective introducing causality filtering to select samples for which contextual information is helpful. To show the effective ness of Causal-CoG we run extensive experiments on 10 multimodal benchmarks and showconsistent improvements e.g. +6.30% on POPE +13.69% on Vizwiz and +6.43% on VQAv2 compared to direct decoding surpassing exist ing methods. We hope Casual-CoG inspires explorations of context knowledge in multimodal models and serves as a plug-and-play strategy for MLM decoding.
****
Hide in Thicket: Generating Imperceptible and Rational Adversarial Perturbations on 3D Point Clouds
Tianrui Lou, Xiaojun Jia, Jindong Gu, Li Liu, Siyuan Liang, Bangyan He, Xiaochun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24326-24335
Adversarial attack methods based on point manipulation for 3D point cloud classification have revealed the fragility of 3D models yet the adversarial examples they produce are easily perceived or defended against. The trade-off between the imperceptibility and adversarial strength leads most point attack methods to inevitably introduce easily detectable outlier points upon a successful attack. Another promising strategy shape-based attack can effectively eliminate outliers but existing methods often suffer significant reductions in imperceptibility due to irrational deformations. We find that concealing deformation perturbations in areas insensitive to human eyes can achieve a better trade-off between imperceptibility and adversarial strength specifically in parts of the object surface that are complex and exhibit drastic curvature changes. Therefore we propose a novel shape-based adversarial attack method HiT-ADV which initially conducts a two-stage search for attack regions based on saliency and imperceptibility scores and then adds deformation perturbations in each attack region using Gaussian kernel functions. Additionally HiT-ADV is extendable to physical attack. We propose that by employing benign resampling and benign rigid transformations we can further enhance physical adversarial strength with little sacrifice to imperceptibility. Extensive experiments have validated the superiority of our method in terms of adversarial and imperceptible properties in both digital and physical spaces.
****
SG-BEV: Satellite-Guided BEV Fusion for Cross-View Semantic Segmentation
Junyan Ye, Qiyan Luo, Jinhua Yu, Huaping Zhong, Zhimeng Zheng, Conghui He, Weijia Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27748-27757
This paper aims at achieving fine-grained building attribute segmentation in a cross-view scenario i.e. using satellite and street-view image pairs. The main challenge lies in overcoming the significant perspective differences between street views and satellite views. In this work we introduce SG-BEV a novel approach for satellite-guided BEV fusion for cross-view semantic segmentation. To overcome the limitations of existing cross-view projection methods in capturing the complete building facade features we innovatively incorporate Bird's Eye View (BEV) method to establish a spatially explicit mapping of street-view features. Moreover we fully leverage the advantages of multiple perspectives by introducing a novel satellite-guided reprojection module optimizing the uneven feature distribut

ion issues associated with traditional BEV methods. Our method demonstrates significant improvements on four cross-view datasets collected from multiple cities including New York San Francisco and Boston. On average across these datasets our method achieves an increase in mIOU by 10.13% and 5.21% compared with the state-of-the-art satellite-based and cross-view methods. The code and datasets of this work will be released at https://github.com/sysu-liweijia-lab/SG-BEV.
****
Brush2Prompt: Contextual Prompt Generator for Object Inpainting
Mang Tik Chiu, Yuqian Zhou, Lingzhi Zhang, Zhe Lin, Connelly Barnes, Sohrab Amirghodsi, Eli Shechtman, Humphrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12636-12645
Object inpainting is a task that involves adding objects to real images and seamlessly compositing them. With the recent commercialization of products like Stable Diffusion and Generative Fill inserting objects into images by using prompts has achieved impressive visual results. In this paper we propose a prompt suggestion model to simplify the process of prompt input. When the user provides an image and a mask our model predicts suitable prompts based on the partial contextual information in the masked image and the shape and location of the mask. Specifically we introduce a concept-diffusion in the CLIP space that predicts CLIP-text embeddings from a masked image. These diffused embeddings can be directly injected into open-source inpainting models like Stable Diffusion and its variants. Alternatively they can be decoded into natural language for use in other publicly available applications such as Generative Fill. Our prompt suggestion model demonstrates a balanced accuracy and diversity showing its capability to be both contextually aware and creatively adaptive.
****
Joint-Task Regularization for Partially Labeled Multi-Task Learning
Kento Nishi, Junsik Kim, Wanhua Li, Hanspeter Pfister; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16152-16162
Multi-task learning has become increasingly popular in the machine learning field but its practicality is hindered by the need for large labeled datasets. Most multi-task learning methods depend on fully labeled datasets wherein each input example is accompanied by ground-truth labels for all target tasks. Unfortunately curating such datasets can be prohibitively expensive and impractical especially for dense prediction tasks which require per-pixel labels for each image. With this in mind we propose Joint-Task Regularization (JTR) an intuitive technique which leverages cross-task relations to simultaneously regularize all tasks in a single joint-task latent space to improve learning when data is not fully labeled for all tasks. JTR stands out from existing approaches in that it regularizes all tasks jointly rather than separately in pairs---therefore it achieves linear complexity relative to the number of tasks while previous methods scale quadratically. To demonstrate the validity of our approach we extensively benchmark our method across a wide variety of partially labeled scenarios based on NYU-v2 Cityscapes and Taskonomy.
****
Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification
Bin Yang, Jun Chen, Mang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16870-16879
Unsupervised visible-infrared person re-identification (US-VI-ReID) centers on learning a cross-modality retrieval model without labels reducing the reliance on expensive cross-modality manual annotation. Previous US-VI-ReID works gravitate toward learning cross-modality information with the deep features extracted from the ultimate layer. Nevertheless interfered by the multiple discrepancies solely relying on deep features is insufficient for accurately learning modality-invariant features resulting in negative optimization. The shallow feature from the shallow layers contains nuanced detail information which is critical for effective cross-modality learning but is disregarded regrettably by the existing methods. To address the above issues we design a Shallow-Deep Collaborative Learning

(SDCL) framework based on the transformer with shallow-deep contrastive learning incorporating Collaborative Neighbor Learning (CNL) and Collaborative Ranking Association (CRA) module. Specifically CNL unveils the intrinsic homogeneous and heterogeneous collaboration which are harnessed for neighbor alignment enhancing the robustness in a dynamic manner. Furthermore CRA associates the cross-modality labels with the ranking association between shallow and deep features furnishing valuable supervision for cross-modality learning. Extensive experiments validate the superiority of our method even outperforming certain supervised counterparts.
****
Dancing with Still Images: Video Distillation via Static-Dynamic Disentanglement
Ziyu Wang, Yue Xu, Cewu Lu, Yong-Lu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6296-6304
Recently dataset distillation has paved the way towards efficient machine learning especially for image datasets. However the distillation for videos characterized by an exclusive temporal dimension remains an underexplored domain. In this work we provide the first systematic study of video distillation and introduce a taxonomy to categorize temporal compression. Our investigation reveals that the temporal information is usually not well learned during distillation and the temporal dimension of synthetic data contributes little. The observations motivate our unified framework of disentangling the dynamic and static information in the videos. It first distills the videos into still images as static memory and then compensates the dynamic and motion information with a learnable dynamic memory block. Our method achieves state-of-the-art on video datasets at different scales with notably smaller memory storage budget. Our code is available at https://github.com/yuz1wan/video_distillation.
****
Context-Aware Integration of Language and Visual References for Natural Language Tracking
Yanyan Shao, Shuting He, Qi Ye, Yuchao Feng, Wenhan Luo, Jiming Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19208-19217
Tracking by natural language specification (TNL) aims to consistently localize a target in a video sequence given a linguistic description in the initial frame. Existing methodologies perform language-based and template-based matching for target reasoning separately and merge the matching results from two sources which suffer from tracking drift when language and visual templates miss-align with the dynamic target state and ambiguity in the later merging stage. To tackle the issues we propose a joint multi-modal tracking framework with 1) a prompt modulation module to leverage the complementarity between temporal visual templates and language expressions enabling precise and context-aware appearance and linguistic cues and 2) a unified target decoding module to integrate the multi-modal reference cues and executes the integrated queries on the search image to predict the target location in an end-to-end manner directly. This design ensures spatio-temporal consistency by leveraging historical visual information and introduces an integrated solution generating predictions in a single step. Extensive experiments conducted on TNL2K OTB-Lang LaSOT and RefCOCOg validate the efficacy of our proposed approach. The results demonstrate competitive performance against state-of-the-art methods for both tracking and grounding. Code is available at https://github.com/twotwo2/QueryNLT
****
An Edit Friendly DDPM Noise Space: Inversion and Manipulations
Inbar Huberman-Spiegelglas, Vladimir Kulikov, Tomer Michaeli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12469-12478
Denoising diffusion probabilistic models (DDPMs) employ a sequence of white Gaussian noise samples to generate an image. In analogy with GANs those noise maps could be considered as the latent code associated with the generated image. However this native noise space does not possess a convenient structure and is thus challenging to work with in editing tasks. Here we propose an alternative latent

noise space for DDPM that enables a wide range of editing operations via simple means and present an inversion method for extracting these edit-friendly noise maps for any given image (real or synthetically generated). As opposed to the native DDPM noise space the edit-friendly noise maps do not have a standard normal distribution and are not statistically independent across timesteps. However they allow perfect reconstruction of any desired image and simple transformations on them translate into meaningful manipulations of the output image (e.g. shifting color edits). Moreover in text-conditional models fixing those noise maps while changing the text prompt modifies semantics while retaining structure. We illustrate how this property enables text-based editing of real images via the diverse DDPM sampling scheme (in contrast to the popular non-diverse DDIM inversion). We also show how it can be used within existing diffusion-based editing methods to improve their quality and diversity. The code of the method is attached to this submission.
****
LEAP-VO: Long-term Effective Any Point Tracking for Visual Odometry
Weirong Chen, Le Chen, Rui Wang, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19844-19853
Visual odometry estimates the motion of a moving camera based on visual input. Existing methods mostly focusing on two-view point tracking often ignore the rich temporal context in the image sequence thereby overlooking the global motion patterns and providing no assessment of the full trajectory reliability. These shortcomings hinder performance in scenarios with occlusion dynamic objects and low-texture areas. To address these challenges we present the Long-term Effective Any Point Tracking (LEAP) module. LEAP innovatively combines visual inter-track and temporal cues with mindfully selected anchors for dynamic track estimation. Moreover LEAP's temporal probabilistic formulation integrates distribution updates into a learnable iterative refinement module to reason about point-wise uncertainty. Based on these traits we develop LEAP-VO a robust visual odometry system adept at handling occlusions and dynamic scenes. Our mindful integration showcases a novel practice by employing long-term point tracking as the front-end. Extensive experiments demonstrate that the proposed pipeline significantly outperforms existing baselines across various visual odometry benchmarks.
****
RoDLA: Benchmarking the Robustness of Document Layout Analysis Models
Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15556-15566
Before developing a Document Layout Analysis (DLA) model in real-world applications conducting comprehensive robustness testing is essential. However the robustness of DLA models remains underexplored in the literature. To address this we are the first to introduce a robustness benchmark for DLA models which includes 450K document images of three datasets. To cover realistic corruptions we propose a perturbation taxonomy with 12 common document perturbations with 3 severity levels inspired by real-world document processing. Additionally to better understand document perturbation impacts we propose two metrics Mean Perturbation Effect (mPE) for perturbation assessment and Mean Robustness Degradation (mRD) for robustness evaluation. Furthermore we introduce a self-titled model i.e. Robust Document Layout Analyzer (RoDLA) which improves attention mechanisms to boost extraction of robust features. Experiments on the proposed benchmarks (PubLayNet-P DocLayNet-P and M6Doc-P) demonstrate that RoDLA obtains state-of-the-art mRD scores of 115.7 135.4 and 150.4 respectively. Compared to previous methods RoDLA achieves notable improvements in mAP of +3.8% +7.1% and +12.1% respectively.
****
UniRepLKNet: A Universal Perception Large-Kernel ConvNet for Audio Video Point Cloud Time-Series and Image Recognition
Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5513-5524
Large-kernel convolutional neural networks (ConvNets) have recently received ext

ensive research attention but two unresolved and critical issues demand further investigation. 1) The architectures of existing large-kernel ConvNets largely follow the design principles of conventional ConvNets or transformers while the architectural design for large-kernel ConvNets remains under-addressed. 2) As transformers have dominated multiple modalities it remains to be investigated whether ConvNets also have a strong universal perception ability in domains beyond vision. In this paper we contribute from two aspects. 1) We propose four architectural guidelines for designing large-kernel ConvNets the core of which is to exploit the essential characteristics of large kernels that distinguish them from small kernels - they can see wide without going deep. Following such guidelines our proposed large-kernel ConvNet shows leading performance in image recognition (ImageNet accuracy of 88.0% ADE20K mIoU of 55.6% and COCO box AP of 56.4%) demonstrating better performance and higher speed than the recent powerful competitors. 2) We discover large kernels are the key to unlocking the exceptional performance of ConvNets in domains where they were originally not proficient. With certain modality-related preprocessing approaches the proposed model achieves state-of-the-art performance on time-series forecasting and audio recognition tasks even without modality-specific customization to the architecture. All the code and models are publicly available on GitHub and Huggingface.
****
Unveiling the Unknown: Unleashing the Power of Unknown to Known in Open-Set Source-Free Domain Adaptation
Fuli Wan, Han Zhao, Xu Yang, Cheng Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24015-24024
Open-Set Source-Free Domain Adaptation aims to transfer knowledge in realistic scenarios where the target domain has additional unknown classes compared to the limited-access source domain. Due to the absence of information on unknown classes existing methods mainly transfer knowledge of known classes while roughly grouping unknown classes as one attenuating the knowledge transfer and generalization. In contrast this paper advocates that exploring unknown classes can better identify known ones and proposes a domain adaptation model to transfer knowledge on known and unknown classes jointly. Specifically given a source pre-trained model we first introduce an unknown diffuser that can determine whether classes in space need to be split and merged through similarity measures to estimate and generate a wider class space distribution including known and unknown classes. Based on such a wider space distribution we enhance the reliability of known class knowledge in the source pre-trained model through contrastive constraint. Finally various supervision information including reliable known class knowledge and clustered pseudo-labels optimize the model for impressive knowledge transfer and generalization. Extensive experiments show that our network can achieve superior exploration and knowledge generalization on unknown classes while with excellent known class transfer. The code is available at https://github.com/xdwfl/UPUK.
****
BilevelPruning: Unified Dynamic and Static Channel Pruning for Convolutional Neural Networks
Shangqian Gao, Yanfu Zhang, Feihu Huang, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16090-16100
Most existing dynamic or runtime channel pruning methods have to store all weights to achieve efficient inference which brings extra storage costs. Static pruning methods can reduce storage costs directly but their performance is limited by using a fixed sub-network to approximate the original model. Most existing pruning works suffer from these drawbacks because they were designed to only conduct either static or dynamic pruning. In this paper we propose a novel method to solve both efficiency and storage challenges via simultaneously conducting dynamic and static channel pruning for convolutional neural networks. We propose a new bi-level optimization based model to naturally integrate the static and dynamic channel pruning. By doing so our method enjoys benefits from both sides and the disadvantages of dynamic and static pruning are reduced. After pruning we permanently remove redundant parameters and then finetune the model with dynamic flexi

bility. Experimental results on CIFAR-10 and ImageNet datasets suggest that our method can achieve state-of-the-art performance compared to existing dynamic and static channel pruning methods.
****
IDGuard: Robust General Identity-centric POI Proactive Defense Against Face Editing Abuse
Yunshu Dai, Jianwei Fei, Fangjun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11934-11943
In this work we propose IDGuard a novel proactive defense method from the perspective of developers to protect Persons-of-Interest (POI) such as national leaders from face editing abuse. We build a bridge between identities and model behavior safeguarding POI identities rather than merely certain face images. Given a face editing model IDGuard enables it to reject editing any image containing POI identities while retaining its editing functionality for regular use. Specifically we insert an ID Normalization Layer into the original face editing model and introduce an ID Extractor to extract the identities of input images. To differentiate the editing behavior between POI and nonPOI we use a transformer-based ID Encoder to encode extracted POI identities as parameters of the ID Normalization Layer. Our method supports the simultaneous protection of multiple POI and allows for the addition of new POI in the inference stage without the need for retraining. Extensive experiments show that our method achieves 100% protection accuracy on POI images even if they are neither included in the training set nor subject to any preprocessing. Notably our method exhibits excellent robustness against image and model attacks and maintains 100% protection performance when generalized to various face editing models further demonstrating its practicality.
****
SwiftBrush: One-Step Text-to-Image Diffusion Model with Variational Score Distillation
Thuan Hoang Nguyen, Anh Tran; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7807-7816
Despite their ability to generate high-resolution and diverse images from text prompts text-to-image diffusion models often suffer from slow iterative sampling processes. Model distillation is one of the most effective directions to accelerate these models. However previous distillation methods fail to retain the generation quality while requiring a significant amount of images for training either from real data or synthetically generated by the teacher model. In response to this limitation we present a novel image-free distillation scheme named SwiftBrush. Drawing inspiration from text-to-3D synthesis in which a 3D neural radiance field that aligns with the input prompt can be obtained from a 2D text-to-image diffusion prior via a specialized loss without the use of any 3D data ground-truth our approach re-purposes that same loss for distilling a pretrained multi-step text-to-image model to a student network that can generate high-fidelity images with just a single inference step. In spite of its simplicity our model stands as one of the first one-step text-to-image generators that can produce images of comparable quality to Stable Diffusion without reliance on any training image data. Remarkably SwiftBrush achieves an FID score of 16.67 and a CLIP score of 0.29 on the COCO-30K benchmark achieving competitive results or even substantially surpassing existing state-of-the-art distillation techniques.
****
DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations
Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8693-8702
The diffusion-based text-to-image model harbors immense potential in transferring reference style. However current encoder-based approaches significantly impair the text controllability of text-to-image models while transferring styles. In this paper we introduce DEADiff to address this issue using the following two strategies: 1) a mechanism to decouple the style and semantics of reference images. The decoupled feature representations are first extracted by Q-Formers which a

re instructed by different text descriptions. Then they are injected into mutually exclusive subsets of cross-attention layers for better disentanglement. 2) A non-reconstructive learning method. The Q-Formers are trained using paired images rather than the identical target in which the reference image and the ground-truth image are with the same style or semantics. We show that DEADiff attains the best visual stylization results and optimal balance between the text controllability inherent in the text-to-image model and style similarity to the reference image as demonstrated both quantitatively and qualitatively. Our project page is https://tianhao-qi.github.io/DEADiff/.
****
Instance-Adaptive and Geometric-Aware Keypoint Learning for Category-Level 6D Object Pose Estimation
Xiao Lin, Wenfei Yang, Yuan Gao, Tianzhu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21040-21049
Category-level 6D object pose estimation aims to estimate the rotation translation and size of unseen instances within specific categories. In this area dense correspondence-based methods have achieved leading performance. However they do not explicitly consider the local and global geometric information of different instances resulting in poor generalization ability to unseen instances with significant shape variations. To deal with this problem we propose a novel Instance-adaptive and To deal with this problem we propose a novel Instance-Adaptive and Geometric-Aware Keypoint Learning method for category-level 6D object pose estimation (AG-Pose) which includes two key designs: (1) The first design is an Instance-Adaptive Keypoint Detection module which can adaptively detect a set of sparse keypoints for various instances to represent their geometric structures. (2) The second design is a Geometric-Aware Feature Aggregation module which can efficiently integrate the local and global geometric information into keypoint features. These two modules can work together to establish robust keypoint-level correspondences for unseen instances thus enhancing the generalization ability of the model.Experimental results on CAMERA25 and REAL275 datasets show that the proposed AG-Pose outperforms state-of-the-art methods by a large margin without category-specific shape priors.
****
Universal Semi-Supervised Domain Adaptation by Mitigating Common-Class Bias
Wenyu Zhang, Qingmu Liu, Felix Ong Wei Cong, Mohamed Ragab, Chuan-Sheng Foo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23912-23921
Domain adaptation is a critical task in machine learning that aims to improve model performance on a target domain by leveraging knowledge from a related source domain. In this work we introduce Universal Semi-Supervised Domain Adaptation (UniSSDA) a practical yet challenging setting where the target domain is partially labeled and the source and target label space may not strictly match. UniSSDA is at the intersection of Universal Domain Adaptation (UniDA) and Semi-Supervised Domain Adaptation (SSDA): the UniDA setting does not allow for fine-grained categorization of target private classes not represented in the source domain while SSDA focuses on the restricted closed-set setting where source and target label spaces match exactly. Existing UniDA and SSDA methods are susceptible to common-class bias in UniSSDA settings where models overfit to data distributions of classes common to both domains at the expense of private classes. We propose a new prior-guided pseudo-label refinement strategy to reduce the reinforcement of common-class bias due to pseudo-labeling a common label propagation strategy in domain adaptation. We demonstrate the effectiveness of the proposed strategy on benchmark datasets Office-Home DomainNet and VisDA. The proposed strategy attains the best performance across UniSSDA adaptation settings and establishes a new baseline for UniSSDA.
****
Exact Fusion via Feature Distribution Matching for Few-shot Image Generation
Yingbo Zhou, Yutong Ye, Pengyu Zhang, Xian Wei, Mingsong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8383-8392

Few-shot image generation as an important yet challenging visual task still suff
ers from the trade-off between generation quality and diversity. According to th
e principle of feature-matching learning existing fusion-based methods usually f
use different features by using similarity measurements or attention mechanisms
which may match features inaccurately and lead to artifacts in the texture and s
tructure of generated images. In this paper we propose an exact Fusion via Featu
re Distribution matching Generative Adversarial Network (F2DGAN) for few-shot im
age generation. The rationale behind this is that feature distribution matching
is much more reliable than feature matching to explore the statistical character
s in image feature space for limited real-world data. To model feature distribut
ions from only a few examples for feature fusion we design a novel variational f
eature distribution matching fusion module to perform exact fusion by empirical
cumulative distribution functions. Specifically we employ a variational autoenco
der to transform deep image features into distributions and fuse different featu
res exactly by applying histogram matching. Additionally we formulate two effect
ive losses to guide the matching process for better fitting our fusion strategy.
 Extensive experiments compared with state-of-the-art methods on three public da
tasets demonstrate the superiority of F2DGAN for few-shot image generation in te
rms of generation quality and diversity and the effectiveness of data augmentati
on in downstream classification tasks.
****
CoDeF: Content Deformation Fields for Temporally Consistent Video Processing
Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xia
owei Zhou, Qifeng Chen, Yujun Shen; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2024, pp. 8089-8099
We present the content deformation field (CoDeF) as a new type of video represen
tation which consists of a canonical content field aggregating the static conten
ts in the entire video and a temporal deformation field recording the transforma
tions from the canonical image (i.e. rendered from the canonical content field)
to each individual frame along the time axis. Given a target video these two fie
lds are jointly optimized to reconstruct it through a carefully tailored renderi
ng pipeline. We advisedly introduce some regularizations into the optimization p
rocess urging the canonical content field to inherit semantics (e.g. the object
shape) from the video. With such a design CoDeF naturally supports lifting image
 algorithms for video processing in the sense that one can apply an image algori
thm to the canonical image and effortlessly propagate the outcomes to the entire
 video with the aid of the temporal deformation field. We experimentally show th
at CoDeF is able to lift image-to-image translation to video-to-video translatio
n and lift keypoint detection to keypoint tracking without any training. More im
portantly thanks to our lifting strategy that deploys the algorithms on only one
 image we achieve superior cross-frame consistency in processed videos compared
to existing video-to-video translation approaches and even manage to track non-r
igid objects like water and smog. Code will be made publicly available.
****
QUADify: Extracting Meshes with Pixel-level Details and Materials from Images
Maximilian Frühauf, Hayko Riemenschneider, Markus Gross, Christopher Schroers; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 4661-4670
Despite exciting progress in automatic 3D reconstruction from images excessive a
nd irregular triangular faces in the resulting meshes still constitute a signifi
cant challenge when it comes to adoption in practical artist workflows. Therefor
e we propose a method to extract regular quad-dominant meshes from posed images.
 More specifically we generate a high-quality 3D model through decomposition int
o an easily editable quad-dominant mesh with pixel-level details such as displac
ement materials and lighting. To enable end-to-end learning of shape and quad to
pology we QUADify a neural implicit representation using our novel differentiabl
e re-meshing objective. Distinct from previous work our method exploits artifact
-free Catmull-Clark subdivision combined with vertex displacement to extract pix
el-level details linked to the base geometry. Finally we apply differentiable re
ndering techniques for material and lighting decomposition to optimize for image

reconstruction. Our experiments show the benefits of end-to-end re-meshing and that our method yields state-of-the-art geometric accuracy while providing light weight meshes with displacements and textures that are directly compatible with professional renderers and game engines.
****
RecDiffusion: Rectangling for Image Stitching with Diffusion Models
Tianhao Zhou, Haipeng Li, Ziyi Wang, Ao Luo, Chen-Lin Zhang, Jiajun Li, Bing Zeng, Shuaicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2692-2701
Image stitching from different captures often results in non-rectangular boundaries which is often considered unappealing. To solve non-rectangular boundaries current solutions involve cropping which discards image content inpainting which can introduce unrelated content or warping which can distort non-linear features and introduce artifacts. To overcome these issues we introduce a novel diffusion-based learning framework RecDiffusion for image stitching rectangling. This framework combines Motion Diffusion Models (MDM) to generate motion fields effectively transitioning from the stitched image's irregular borders to a geometrically corrected intermediary. Followed by Content Diffusion Models (CDM) for image detail refinement. Notably our sampling process utilizes a weighted map to identify regions needing correction during each iteration of CDM. Our RecDiffusion ensures geometric accuracy and overall visual appeal surpassing all previous methods in both quantitative and qualitative measures when evaluated on public benchmarks. Code is released at https://github.com/lhaippp/RecDiffusion.
****
Eclipse: Disambiguating Illumination and Materials using Unintended Shadows
Dor Verbin, Ben Mildenhall, Peter Hedman, Jonathan T. Barron, Todd Zickler, Pratul P. Srinivasan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 77-86
Decomposing an object's appearance into representations of its materials and the surrounding illumination is difficult even when the object's 3D shape is known beforehand. This problem is especially challenging for diffuse objects: it is ill-conditioned because diffuse materials severely blur incoming light and it is ill-posed because diffuse materials under high-frequency lighting can be indistinguishable from shiny materials under low-frequency lighting. We show that it is possible to recover precise materials and illumination---even from diffuse objects---by exploiting unintended shadows like the ones cast onto an object by the photographer who moves around it. These shadows are a nuisance in most previous inverse rendering pipelines but here we exploit them as signals that improve conditioning and help resolve material-lighting ambiguities. We present a method based on differentiable Monte Carlo ray tracing that uses images of an object to jointly recover its spatially-varying materials the surrounding illumination environment and the shapes of the unseen light occluders who inadvertently cast shadows upon it.
****
Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields
Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, Achuta Kadambi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 21676-21685
3D scene representations have gained immense popularity in recent years. Methods that use Neural Radiance fields are versatile for traditional tasks such as novel view synthesis. In recent times some work has emerged that aims to extend the functionality of NeRF beyond view synthesis for semantically aware tasks such as editing and segmentation using 3D feature field distillation from 2D foundation models. However these methods have two major limitations: (a) they are limited by the rendering speed of NeRF pipelines and (b) implicitly represented feature fields suffer from continuity artifacts reducing feature quality. Recently 3D Gaussian Splatting has shown state-of-the-art performance on real-time radiance field rendering. In this work we go one step further: in addition to radiance fie

ld rendering we enable 3D Gaussian splatting on arbitrary-dimension semantic features via 2D foundation model distillation. This translation is not straightforward: naively incorporating feature fields in the 3DGS framework encounters significant challenges notably the disparities in spatial resolution and channel consistency between RGB images and feature maps. We propose architectural and training changes to efficiently avert this problem. Our proposed method is general and our experiments showcase novel view semantic segmentation language-guided editing and segment anything through learning feature fields from state-of-the-art 2D foundation models such as SAM and CLIP-LSeg. Across experiments our distillation method is able to provide comparable or better results while being significantly faster to both train and render. Additionally to the best of our knowledge we are the first method to enable point and bounding-box prompting for radiance field manipulation by leveraging the SAM model. Project website at: https://feature-3dgs.github.io/
****
Balancing Act: Distribution-Guided Debiasing in Diffusion Models
Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6668-6678
Diffusion Models (DMs) have emerged as powerful generative models with unprecedented image generation capability. These models are widely used for data augmentation and creative applications. However DMs reflect the biases present in the training datasets. This is especially concerning in the context of faces where the DM prefers one demographic subgroup vs others (eg. female vs male). In this work we present a method for debiasing DMs without relying on additional reference data or model retraining. Specifically we propose Distribution Guidance which enforces the generated images to follow the prescribed attribute distribution. To realize this we build on the key insight that the latent features of denoising UNet hold rich demographic semantics and the same can be leveraged to guide debiased generation. We train Attribute Distribution Predictor (ADP) - a small mlp that maps the latent features to the distribution of attributes. ADP is trained with pseudo labels generated from existing attribute classifiers. The proposed Distribution Guidance with ADP enables us to do fair generation. Our method reduces bias across single/multiple attributes and outperforms the baseline by a significant margin for unconditional and text-conditional diffusion models. Further we present a downstream task of training a fair attribute classifier by augmenting the training set with our generated data.
****
Viewpoint-Aware Visual Grounding in 3D Scenes
Xiangxi Shi, Zhonghua Wu, Stefan Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14056-14065
Referring expressions for visual objects often include descriptions of relative spatial arrangements to other objects -- e.g. "to the right of" -- that depend on the point of view of the speaker. In 2D referring expression tasks this viewpoint is captured unambiguously in the image. However grounding expressions with such spatial language in 3D without viewpoint annotations can be ambiguous. In this paper we investigate the significance of viewpoint information in 3D visual grounding -- introducing a model that explicitly predicts the speaker's viewpoint based on the referring expression and scene. We pretrain this model on a synthetically generated dataset that provides viewpoint annotations and then finetune on 3D referring expression datasets. Further we introduce an auxiliary uniform object representation loss to encourage viewpoint invariance in learned object representations. We find that our proposed ViewPoint Prediction Network (VPP-Net) achieves state-of-the-art performance on ScanRefer SR3D and NR3D -- improving Accuracy@0.25IoU by 1.06% 0.60% and 2.00% respectively compared to prior work.
****
4K4D: Real-Time 4D View Synthesis at 4K Resolution
Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20029-20040

This paper targets high-fidelity and real-time view synthesis of dynamic 3D scen
es at 4K resolution. Recent methods on dynamic view synthesis have shown impress
ive rendering quality. However their speed is still limited when rendering high-
resolution images. To overcome this problem we propose 4K4D a 4D point cloud rep
resentation that supports hardware rasterization and network pre-computation to
enable unprecedented rendering speed with a high rendering quality. Our represen
tation is built on a 4D feature grid so that the points are naturally regularize
d and can be robustly optimized. In addition we design a novel hybrid appearance
 model that significantly boosts the rendering quality while preserving efficien
cy. Moreover we develop a differentiable depth peeling algorithm to effectively
learn the proposed model from RGB videos. Experiments show that our representati
on can be rendered at over 400 FPS on the DNA-Rendering dataset at 1080p resolut
ion and 80 FPS on the ENeRF-Outdoor dataset at 4K resolution using an RTX 4090 G
PU which is 30x faster than previous methods and achieves the state-of-the-art r
endering quality. Our project page is available at https://zju3dv.github.io/4k4d
.
****
View-decoupled Transformer for Person Re-identification under Aerial-ground Came
ra Network
Quan Zhang, Lei Wang, Vishal M. Patel, Xiaohua Xie, Jianhaung Lai; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
4, pp. 22000-22009
Existing person re-identification methods have achieved remarkable advances in a
ppearance-based identity association across homogeneous cameras such as ground-g
round matching. However as a more practical scenario aerial-ground person re-ide
ntification (AGPReID) among heterogeneous cameras has received minimal attention
. To alleviate the disruption of discriminative identity representation by drama
tic view discrepancy as the most significant challenge in AGPReID the view-decou
pled transformer (VDT) is proposed as a simple yet effective framework. Two majo
r components are designed in VDT to decouple view-related and view-unrelated fea
tures namely hierarchical subtractive separation and orthogonal loss where the f
ormer separates these two features inside the VDT and the latter constrains thes
e two to be independent. In addition we contribute a large-scale AGPReID dataset
 called CARGO consisting of five/eight aerial/ground cameras 5000 identities and
 108563 images. Experiments on two datasets show that VDT is a feasible and effe
ctive solution for AGPReID surpassing the previous method on mAP/Rank1 by up to
5.0%/2.7% on CARGO and 3.7%/5.2% on AG-ReID keeping the same magnitude of comput
ational complexity. Our project is available at https://github.com/LinlyAC/VDT-A
GPReID
****
CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Disti
llation
Lingjun Zhao, Jingyu Song, Katherine A. Skinner; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15470-15480
In the field of 3D object detection for autonomous driving LiDAR-Camera (LC) fus
ion is the top-performing sensor configuration. Still LiDAR is relatively high c
ost which hinders adoption of this technology for consumer automobiles. Alternat
ively camera and radar are commonly deployed on vehicles already on the road tod
ay but performance of Camera-Radar (CR) fusion falls behind LC fusion. In this w
ork we propose Camera-Radar Knowledge Distillation (CRKD) to bridge the performa
nce gap between LC and CR detectors with a novel cross-modality KD framework. We
 use the Bird's-Eye-View (BEV) representation as the shared feature space to ena
ble effective knowledge distillation. To accommodate the unique cross-modality K
D path we propose four distillation losses to help the student learn crucial fea
tures from the teacher model. We present extensive evaluations on the nuScenes d
ataset to demonstrate the effectiveness of the proposed CRKD framework. The proj
ect page for CRKD is https://song-jingyu.github.io/CRKD.
****
Differentiable Point-based Inverse Rendering
Hoon-Gyu Chung, Seokjun Choi, Seung-Hwan Baek; Proceedings of the IEEE/CVF Confe

rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4399-4409
We present differentiable point-based inverse rendering DPIR an analysis-by-synthesis method that processes images captured under diverse illuminations to estimate shape and spatially-varying BRDF. To this end we adopt point-based rendering eliminating the need for multiple samplings per ray typical of volumetric rendering thus significantly enhancing the speed of inverse rendering. To realize this idea we devise a hybrid point-volumetric representation for geometry and a regularized basis-BRDF representation for reflectance. The hybrid geometric representation enables fast rendering through point-based splatting while retaining the geometric details and stability inherent to SDF-based representations. The regularized basis-BRDF mitigates the ill-posedness of inverse rendering stemming from limited light-view angular samples. We also propose an efficient shadow detection method using point-based shadow map rendering. Our extensive evaluations demonstrate that DPIR outperforms prior works in terms of reconstruction accuracy computational efficiency and memory footprint. Furthermore our explicit point-based representation and rendering enables intuitive geometry and reflectance editing.
****
OED: Towards One-stage End-to-End Dynamic Scene Graph Generation
Guan Wang, Zhimin Li, Qingchao Chen, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27938-27947
Dynamic Scene Graph Generation (DSGG) focuses on identifying visual relationships within the spatial-temporal domain of videos. Conventional approaches often employ multi-stage pipelines which typically consist of object detection temporal association and multi-relation classification. However these methods exhibit inherent limitations due to the separation of multiple stages and independent optimization of these sub-problems may yield sub-optimal solutions. To remedy these limitations we propose a one-stage end-to-end framework termed OED which streamlines the DSGG pipeline. This framework reformulates the task as a set prediction problem and leverages pair-wise features to represent each subject-object pair within the scene graph. Moreover another challenge of DSGG is capturing temporal dependencies we introduce a Progressively Refined Module (PRM) for aggregating temporal context without the constraints of additional trackers or handcrafted trajectories enabling end-to-end optimization of the network. Extensive experiments conducted on the Action Genome benchmark demonstrate the effectiveness of our design. The code and models are available at https://github.com/guanw-pku/OED.
****
CoG-DQA: Chain-of-Guiding Learning with Large Language Models for Diagram Question Answering
Shaowei Wang, Lingling Zhang, Longji Zhu, Tao Qin, Kim-Hui Yap, Xinyu Zhang, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13969-13979
Diagram Question Answering (DQA) is a challenging task requiring models to answer natural language questions based on visual diagram contexts. It serves as a crucial basis for academic tutoring technical support and more practical applications. DQA poses significant challenges such as the demand for domain-specific knowledge and the scarcity of annotated data which restrict the applicability of large-scale deep models. Previous approaches have explored external knowledge integration through pre-training but these methods are costly and can be limited by domain disparities. While Large Language Models (LLMs) show promise in question-answering there is still a gap in how to cooperate and interact with the diagram parsing process. In this paper we introduce the Chain-of-Guiding Learning Model for Diagram Question Answering (CoG-DQA) a novel framework that effectively addresses DQA challenges. CoG-DQA leverages LLMs to guide diagram parsing tools (DPTs) through the guiding chains enhancing the precision of diagram parsing while introducing rich background knowledge. Our experimental findings reveal that CoG-DQA surpasses all comparison models in various DQA scenarios achieving an average accuracy enhancement exceeding 5% and peaking at 11% across four datasets. These results underscore CoG-DQA's capacity to advance the field of visual question answering and promote the integration of LLMs into specialized domains.

****
Transferable and Principled Efficiency for Open-Vocabulary Segmentation
Jingxuan Xu, Wuyang Chen, Yao Zhao, Yunchao Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15814-15824
Recent success of pre-trained foundation vision-language models makes Open-Vocabulary Segmentation (OVS) possible. Despite the promising performance this approach introduces heavy computational overheads for two challenges: 1) large model sizes of the backbone; 2) expensive costs during the fine-tuning. These challenges hinder this OVS strategy from being widely applicable and affordable in real-world scenarios. Although traditional methods such as model compression and efficient fine-tuning can address these challenges they often rely on heuristics. This means that their solutions cannot be easily transferred and necessitate re-training on different models which comes at a cost. In the context of efficient OVS we target achieving performance that is comparable to or even better than prior OVS works based on large vision-language foundation models by utilizing smaller models that incur lower training costs. The core strategy is to make our efficiency principled and thus seamlessly transferable from one OVS framework to others without further customization. Comprehensive experiments on diverse OVS benchmarks demonstrate our superior trade-off between segmentation accuracy and computation costs over previous works. Our code is available on https://github.com/Xujxyang/OpenTrans
****
A Unified and Interpretable Emotion Representation and Expression Generation
Reni Paskaleva, Mykyta Holubakha, Andela Ilic, Saman Motamed, Luc Van Gool, Danda Paudel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2447-2456
Canonical emotions such as happy sad and fear are easy to understand and annotate. However emotions are often compound e.g. happily surprised and can be mapped to the action units (AUs) used for expressing emotions and trivially to the canonical ones. Intuitively emotions are continuous as represented by the arousal-valence (AV) model. An interpretable unification of these four modalities --namely Canonical Compound AUs and AV-- is highly desirable for a better representation and understanding of emotions. However such unification remains to be unknown in the current literature. In this work we propose an interpretable and unified emotion model referred as C2A2. We also develop a method that leverages labels of the non-unified models to annotate the novel unified one. Finally we modify the text-conditional diffusion models to understand continuous numbers which are then used to generate continuous expressions using our unified emotion model. Through quantitative and qualitative experiments we show that our generated images are rich and capture subtle expressions. Our work allows a fine-grained generation of expressions in conjunction with other textual inputs and offers a new label space for emotions at the same time.
****
Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution
Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2535-2545
Text-based diffusion models have exhibited remarkable success in generation and editing showing great promise for enhancing visual content with their generative prior. However applying these models to video super-resolution remains challenging due to the high demands for output fidelity and temporal consistency which is complicated by the inherent randomness in diffusion models. Our study introduces Upscale-A-Video a text-guided latent diffusion framework for video upscaling. This framework ensures temporal coherence through two key mechanisms: locally it integrates temporal layers into U-Net and VAE-Decoder maintaining consistency within short sequences; globally without training a flow-guided recurrent latent propagation module is introduced to enhance overall video stability by propagating and fusing latent across the entire sequences. Thanks to the diffusion paradigm our model also offers greater flexibility by allowing text prompts to guide

texture creation and adjustable noise levels to balance restoration and generati on enabling a trade-off between fidelity and quality. Extensive experiments show that Upscale-A-Video surpasses existing methods in both synthetic and real-worl d benchmarks as well as in AI-generated videos showcasing impressive visual real ism and temporal consistency.
****
EvDiG: Event-guided Direct and Global Components Separation
Xinyu Zhou, Peiqi Duan, Boyu Li, Chu Zhou, Chao Xu, Boxin Shi; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p p. 9612-9621
Separating the direct and global components of a scene aids in shape recovery an d basic material understanding. Conventional methods capture multiple frames und er high frequency illumination patterns or shadows requiring the scene to keep s tationary during the image acquisition process. Single-frame methods simplify th e capture procedure but yield lower-quality separation results. In this paper we leverage the event camera to facilitate the separation of direct and global com ponents enabling video-rate separation of high quality. In detail we adopt an ev ent camera to record rapid illumination changes caused by the shadow of a line o ccluder sweeping over the scene and reconstruct the coarse separation results th rough event accumulation. We then design a network to resolve the noise in the c oarse separation results and restore color information. A real-world dataset is collected using a hybrid camera system for network training and evaluation. Expe rimental results show superior performance over state-of-the-art methods.
****
DeIL: Direct-and-Inverse CLIP for Open-World Few-Shot Learning
Shuai Shao, Yu Bai, Yan Wang, Baodi Liu, Yicong Zhou; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28505-28514
Open-World Few-Shot Learning (OFSL) is a critical field of research concentratin g on the precise identification of target samples in environments with scarce da ta and unreliable labels thus possessing substantial practical significance. Rec ently the evolution of foundation models like CLIP has revealed their strong cap acity for representation even in settings with restricted resources and data. Th is development has led to a significant shift in focus transitioning from the tr aditional method of "building models from scratch" to a strategy centered on "ef ficiently utilizing the capabilities of foundation models to extract relevant pr ior knowledge tailored for OFSL and apply it judiciously". Amidst this backdrop we unveil the Direct-and-Inverse CLIP (DeIL) an innovative method leveraging our proposed "Direct-and-Inverse" concept to activate CLIP-based methods for addres sing OFSL. This concept transforms conventional single-step classification into a nuanced two-stage process: initially filtering out less probable categories fo llowed by accurately determining the specific category of samples. DeIL comprise s two key components: a pre-trainer (frozen) for data denoising and an adapter (tunable) for achieving precise final classification. In experiments DeIL achieve s SOTA performance on 11 datasets.
****
4D-DRESS: A 4D Dataset of Real-World Human Clothing With Semantic Annotations
Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan J ose Zarate, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2024, pp. 550-560
The studies of human clothing for digital avatars have predominantly relied on s ynthetic datasets. While easy to collect synthetic data often fall short in real ism and fail to capture authentic clothing dynamics. Addressing this gap we intr oduce 4D-DRESS the first real-world 4D dataset advancing human clothing research with its high-quality 4D textured scans and garment meshes. 4D-DRESS captures 6 4 outfits in 520 human motion sequences amounting to 78k textured scans. Creatin g a real-world clothing dataset is challenging particularly in annotating and se gmenting the extensive and complex 4D human scans. To address this we develop a semi-automatic 4D human parsing pipeline. We efficiently combine a human-in-the-loop process with automation to accurately label 4D scans in diverse garments an

d body movements. Leveraging precise annotations and high-quality garment meshes we establish several benchmarks for clothing simulation and reconstruction. 4D-DRESS offers realistic and challenging data that complements synthetic sources paving the way for advancements in research of lifelike human clothing. Website: https://ait.ethz.ch/4d-dress
****
Feedback-Guided Autonomous Driving
Jimuyang Zhang, Zanming Huang, Arijit Ray, Eshed Ohn-Bar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15000-15011
While behavior cloning has recently emerged as a highly successful paradigm for autonomous driving humans rarely learn to perform complex tasks such as driving via imitation or behavior cloning alone. In contrast learning in humans often involves additional detailed guidance throughout the interactive learning process i.e. where feedback often via language provides detailed information as to which part of their trial was performed incorrectly or suboptimally and why. Motivated by this observation we introduce an efficient feedback-based framework for improving behavior-cloning-based training of sensorimotor driving agents. Our key insight is to leverage recent advances in Large Language Models (LLMs) to provide corrective fine-grained feedback regarding the underlying reason behind driving prediction failures. Moreover our introduced network architecture is efficient enabling the first sensorimotor end-to-end training and evaluation of LLM-based driving models. The resulting agent achieves state-of-the-art performance in open-loop evaluation on nuScenes outperforming prior state-of-the-art by over 8.1% and 57.1% in accuracy and collision rate respectively. In CARLA our camera-based agent improves by 16.6% in driving score over prior LIDAR-based approaches.
****
Large Language Models are Good Prompt Learners for Low-Shot Image Classification
Zhaoheng Zheng, Jingmin Wei, Xuefeng Hu, Haidong Zhu, Ram Nevatia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28453-28462
Low-shot image classification where training images are limited or inaccessible has benefited from recent progress on pre-trained vision-language (VL) models with strong generalizability e.g. CLIP. Prompt learning methods built with VL models generate text features from the class names that only have confined class-specific information. Large Language Models (LLMs) with their vast encyclopedic knowledge emerge as the complement. Thus in this paper we discuss the integration of LLMs to enhance pre-trained VL models specifically on low-shot classification. However the domain gap between language and vision blocks the direct application of LLMs. Thus we propose LLaMP Large Language Models as Prompt learners that produces adaptive prompts for the CLIP text encoder establishing it as the connecting bridge. Experiments show that compared with other state-of-the-art prompt learning methods LLaMP yields better performance on both zero-shot generalization and few-shot image classification over a spectrum of 11 datasets. Code will be made available at: https://github.com/zhaohengz/LLaMP.
****
Specularity Factorization for Low-Light Enhancement
Saurabh Saini, P J Narayanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1-12
We present a new additive image factorization technique that treats images to be composed of multiple latent specular components which can be simply estimated recursively by modulating the sparsity during decomposition. Our model-driven RSFNet estimates these factors by unrolling the optimization into network layers requiring only a few scalars to be learned. The resultant factors are interpretable by design and can be fused for different image enhancement tasks via a network or combined directly by the user in a controllable fashion. Based on RSFNet we detail a zero-reference Low Light Enhancement (LLE) application trained without paired or unpaired supervision. Our system improves the state-of-the-art performance on standard benchmarks and achieves better generalization on multiple other datasets. We also integrate our factors with other task specific fusion network

s for applications like deraining deblurring and dehazing with negligible overhe
ad thereby highlighting the multi-domain and multi-task generalizability of our proposed RSFNet. The code and data is released for reproducibility on the projec
t homepage.
****
Paint3D: Paint Anything 3D with Lighting-Less Texture Diffusion Models
Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, Gang Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4252-4262
This paper presents Paint3D a novel coarse-to-fine generative framework that is capable of producing high-resolution lighting-less and diverse 2K UV texture maps for untextured 3D meshes conditioned on text or image inputs. The key challenge addressed is generating high-quality textures without embedded illumination information which allows the textures to be re-lighted or re-edited within modern graphics pipelines. To achieve this our method first leverages a pre-trained depth-aware 2D diffusion model to generate view-conditional images and perform multi-view texture fusion producing an initial coarse texture map. However as 2D models cannot fully represent 3D shapes and disable lighting effects the coarse texture map exhibits incomplete areas and illumination artifacts. To resolve this we train separate UV Inpainting and UVHD diffusion models specialized for the shape-aware refinement of incomplete areas and the removal of illumination artifacts. Through this coarse-to-fine process Paint3D can produce high-quality 2K UV textures that maintain semantic consistency while being lighting-less significantly advancing the state-of-the-art in texturing 3D objects.
****
VILA: On Pre-training for Visual Language Models
Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, Song Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26689-26699
Visual language models (VLMs) rapidly progressed with the recent success of large language models. There have been growing efforts on visual instruction tuning to extend the LLM with visual inputs but lacks an in-depth study of the visual language pre-training process where the model learns to perform joint modeling on both modalities. In this work we examine the design options for VLM pre-training by augmenting LLM towards VLM through step-by-step con- trollable comparisons. We introduce three main findings: (1) freezing LLMs during pre-training can achieve decent zero-shot performance but lack in-context learning capabil- ity which requires unfreezing the LLM; (2) interleaved pre- training data is beneficial whereas image-text pairs alone are not optimal; (3) re-blending text-only instruction data to image-text data during instruction fine-tuning not only remedies the degradation of text-only tasks but also boosts VLM task accuracy. With an enhanced pre-training recipe we build VILA a Visual Language model family that consis- tently outperforms the state-of-the-art models e.g. LLaVA- 1.5 across main benchmarks without bells and whistles. Multi-modal pre-training also helps unveil appealing prop- erties of VILA including multi-image reasoning enhanced in-context learning and better world knowledge. VILA is also deployable on Jetson Orin for on-device VLM.
****
DiLiGenRT: A Photometric Stereo Dataset with Quantified Roughness and Translucency
Heng Guo, Jieji Ren, Feishi Wang, Boxin Shi, Mingjun Ren, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11810-11820
Photometric stereo faces challenges from non-Lambertian reflectance in real-world scenarios. Systematically measuring the reliability of photometric stereo methods in handling such complex reflectance necessitates a real-world dataset with quantitatively controlled reflectances. This paper introduces DiLiGenRT the first real-world dataset for evaluating photometric stereo methods under quantified reflectances by manufacturing 54 hemispheres with varying degrees of two reflectance properties: Roughness and Translucency. Unlike qualitative and semantic lab

els such as diffuse and specular that have been used in previous datasets our qu antified dataset allows comprehensive and systematic benchmark evaluations. In a ddition it facilitates selecting best-fit photometric stereo methods based on th e quantitative reflectance properties. Our dataset and benchmark results are ava ilable at https://photometricstereo.github.io/diligentrt.html.
****
De-Diffusion Makes Text a Strong Cross-Modal Interface
Chen Wei, Chenxi Liu, Siyuan Qiao, Zhishuai Zhang, Alan Yuille, Jiahui Yu; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2024, pp. 13492-13503
We demonstrate text as a strong cross-modal interface. Rather than relying on de ep embeddings to connect image and language as the interface representation our approach represents an image as text from which we enjoy the interpretability an d flexibility inherent to natural language. We employ an autoencoder that uses a pre-trained text-to-image diffusion model for decoding. The encoder is trained to transform an input image into text which is then fed into the fixed text-to-i mage diffusion decoder to reconstruct the original input a process we term De-Di ffusion. Experiments validate both the precision and comprehensiveness of De-Dif fusion text representing images such that it can be readily ingested by off-the-shelf text-to-image tools and LLMs for diverse multi-modal tasks. For example a single De-Diffusion model can generalize to provide transferable prompts for dif ferent text-to-image tools and also achieves a new state of the art on open-ende d vision-language tasks by simply prompting large language models with few-shot examples. Project page: https://dediffusion.github.io/
****
End-to-End Spatio-Temporal Action Localisation with Video Transformers
Alexey A. Gritsenko, Xuehan Xiong, Josip Djolonga, Mostafa Dehghani, Chen Sun, M ario Lucic, Cordelia Schmid, Anurag Arnab; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18373-18383
The most performant spatio-temporal action localisation models use external pers on proposals and complex external memory banks. We propose a fully end-to-end tr ansformer based model that directly ingests an input video and outputs tubelets -- a sequence of bounding boxes and the action classes at each frame. Our flexib le model can be trained with either sparse bounding-box supervision on individua l frames or full tubelet annotations. And in both cases it predicts coherent tub elets as the output. Moreover our end-to-end model requires no additional pre-pr ocessing in the form of proposals or post-processing in terms of non-maximal sup pression. We perform extensive ablation experiments and significantly advance th e state-of-the-art on five different spatio-temporal action localisation benchma rks with both sparse keyframes and full tubelet annotations.
****
Text-Guided Variational Image Generation for Industrial Anomaly Detection and Se gmentation
Mingyu Lee, Jongwon Choi; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2024, pp. 26519-26528
We propose a text-guided variational image generation method to address the chal lenge of getting clean data for anomaly detection in industrial manufacturing. O ur method utilizes text information about the target object learned from extensi ve text library documents to generate non-defective data images resembling the i nput image. The proposed framework ensures that the generated non-defective imag es align with anticipated distributions derived from textual and image-based kno wledge ensuring stability and generality. Experimental results demonstrate the e ffectiveness of our approach surpassing previous methods even with limited non-d efective data. Our approach is validated through generalization tests across fou r baseline models and three distinct datasets. We present an additional analysis to enhance the effectiveness of anomaly detection models by utilizing the gener ated images.
****
Self-Adaptive Reality-Guided Diffusion for Artifact-Free Super-Resolution
Qingping Zheng, Ling Zheng, Yuanfan Guo, Ying Li, Songcen Xu, Jiankang Deng, Han

g Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2024, pp. 25806-25816
Artifact-free super-resolution (SR) aims to translate low-resolution images into their high-resolution counterparts with a strict integrity of the original content eliminating any distortions or synthetic details. While traditional diffusion-based SR techniques have demonstrated remarkable abilities to enhance image detail they are prone to artifact introduction during iterative procedures. Such artifacts ranging from trivial noise to unauthentic textures deviate from the true structure of the source image thus challenging the integrity of the super-resolution process. In this work we propose Self-Adaptive Reality-Guided Diffusion (SARGD) a training-free method that delves into the latent space to effectively identify and mitigate the propagation of artifacts. Our SARGD begins by using an artifact detector to identify implausible pixels creating a binary mask that highlights artifacts. Following this the Reality Guidance Refinement (RGR) process refines artifacts by integrating this mask with realistic latent representations improving alignment with the original image. Nonetheless initial realistic-latent representations from lower-quality images result in over-smoothing in the final output. To address this we introduce a Self-Adaptive Guidance (SAG) mechanism. It dynamically computes a reality score enhancing the sharpness of the realistic latent. These alternating mechanisms collectively achieve artifact-free super-resolution. Extensive experiments demonstrate the superiority of our method delivering detailed artifact-free high-resolution images while reducing sampling steps by 2X. We release our code at https://github.com/ProAirVerse/Self-Adaptive-Guidance-Diffusion.git.
****
End-to-End Temporal Action Detection with 1B Parameters Across 1000 Frames
Shuming Liu, Chen-Lin Zhang, Chen Zhao, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18591-18601
Recently temporal action detection (TAD) has seen significant performance improvement with end-to-end training. However due to the memory bottleneck only models with limited scales and limited data volumes can afford end-to-end training which inevitably restricts TAD performance. In this paper we reduce the memory consumption for end-to-end training and manage to scale up the TAD backbone to 1 billion parameters and the input video to 1536 frames leading to significant detection performance. The key to our approach lies in our proposed temporal-informative adapter (TIA) which is a novel lightweight module that reduces training memory. Using TIA we free the humongous backbone from learning to adapt to the TAD task by only updating the parameters in TIA. TIA also leads to better TAD representation by temporally aggregating context from adjacent frames throughout the backbone. We evaluate our model across four representative datasets. Owing to our efficient design we are able to train end-to-end on VideoMAEv2-giant and achieve 75.4% mAP on THUMOS14 being the first end-to-end model to outperform the best feature-based methods.
****
Multimodal Representation Learning by Alternating Unimodal Adaptation
Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, Huaxiu Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27456-27466
Multimodal learning which integrates data from diverse sensory modes plays a pivotal role in artificial intelligence. However existing multimodal learning methods often struggle with challenges where some modalities appear more dominant than others during multimodal learning resulting in suboptimal performance. To address this challenge we propose MLA (Multimodal Learning with Alternating Unimodal Adaptation). MLA reframes the conventional joint multimodal learning process by transforming it into an alternating unimodal learning process thereby minimizing interference between modalities. Simultaneously it captures cross-modal interactions through a shared head which undergoes continuous optimization across different modalities. This optimization process is controlled by a gradient modification mechanism to prevent the shared head from losing previously acquired inform

ation. During the inference phase MLA utilizes a test-time uncertainty-based mod el fusion mechanism to integrate multimodal information. Extensive experiments a re conducted on five diverse datasets encompassing scenarios with complete modal ities and scenarios with missing modalities. These experiments demonstrate the s uperiority of MLA over competing prior approaches. Our code is available at http s://github.com/Cecile-hi/Multimodal-Learning-with-Alternating-Unimodal-Adaptatio n.
****
MS-MANO: Enabling Hand Pose Tracking with Biomechanical Constraints
Pengfei Xie, Wenqiang Xu, Tutian Tang, Zhenjun Yu, Cewu Lu; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2382-2392
This work proposes a novel learning framework for visual hand dynamics analysis that takes into account the physiological aspects of hand motion. The existing m odels which are simplified joint-actuated systems often produce unnatural motion s. To address this we integrate a musculoskeletal system with a learnable parame tric hand model MANO to create a new model MS-MANO. This model emulates the dyna mics of muscles and tendons to drive the skeletal system imposing physiologicall y realistic constraints on the resulting torque trajectories. We further propose a simulation-in-the-loop pose refinement framework BioPR that refines the initi al estimated pose through a multi-layer perceptron (MLP) network. Our evaluation of the accuracy of MS-MANO and the efficacy of the BioPR is conducted in two se parate parts. The accuracy of MS-MANO is compared with MyoSuite while the effica cy of BioPR is benchmarked against two large-scale public datasets and two recen t state-of-the-art methods. The results demonstrate that our approach consistent ly improves the baseline methods both quantitatively and qualitatively.
****
Generate Like Experts: Multi-Stage Font Generation by Incorporating Font Transfe r Process into Diffusion Models
Bin Fu, Fanghua Yu, Anran Liu, Zixuan Wang, Jie Wen, Junjun He, Yu Qiao; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6892-6901
Few-shot font generation (FFG) produces stylized font images with a limited numb er of reference samples which can significantly reduce labor costs in manual fon t designs. Most existing FFG methods follow the style-content disentanglement pa radigm and employ the Generative Adversarial Network (GAN) to generate target fo nts by combining the decoupled content and style representations. The complicate d structure and detailed style are simultaneously generated in those methods whi ch may be the sub-optimal solutions for FFG task. Inspired by most manual font d esign processes of expert designers in this paper we model font generation as a multi-stage generative process. Specifically as the injected noise and the data distribution in diffusion models can be well-separated into different sub-spaces we are able to incorporate the font transfer process into these models. Based o n this observation we generalize diffusion methods to model font generative proc ess by separating the reverse diffusion process into three stages with different functions: The structure construction stage first generates the structure infor mation for the target character based on the source image and the font transfer stage subsequently transforms the source font to the target font. Finally the fo nt refinement stage enhances the appearances and local details of the target fon t images. Based on the above multi-stage generative process we construct our fon t generation framework named MSD-Font with a dual-network approach to generate f ont images. The superior performance demonstrates the effectiveness of our model . The code is available at: https://github.com/fubinfb/MSD-Font .
****
Pre-training Vision Models with Mandelbulb Variations
Benjamin Naoto Chiche, Yuto Horikawa, Ryo Fujita; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22062-2207 1
The use of models that have been pre-trained on natural image datasets like Imag eNet may face some limitations. First this use may be restricted due to copyrigh

t and license on the training images and privacy laws. Second these datasets and models may incorporate societal and ethical biases. Formula-driven supervised learning (FDSL) enables model pre-training to circumvent these issues. This consists of generating a synthetic image dataset based on mathematical formulae and pre-training the model on it. In this work we propose novel FDSL datasets based on Mandelbulb Variations. These datasets contain RGB images that are projections of colored objects deriving from the 3D Mandelbulb fractal. Pre-training ResNet-50 on one of our proposed datasets MandelbulbVAR-1k enables an average top-1 accuracy over target classification datasets that is at least 1% higher than pre-training on existing FDSL datasets. With regard to anomaly detection on MVTec AD pre-training the WideResNet-50 backbone on MandelbulbVAR-1k enables PatchCore to achieve 97.2% average image-level AUROC. This is only 1.9% lower than pre-training on ImageNet-1k (99.1%) and 4.5% higher than pre-training on the second-best performing FDSL dataset i.e. VisualAtom-1k (92.7%). Regarding Vision Transformer (ViT) pre-training another dataset that we propose and coin MandelbulbVAR-Hybrid-21k enables ViT-Base to achieve 82.2% top-1 accuracy on ImageNet-1k which is 0.4% higher than pre-training on ImageNet-21k (81.8%) and only 0.1% lower than pre-training on VisualAtom-1k (82.3%).
****
Diffuse Attend and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion
Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, Mar Gonzalez-Franco; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3554-3563
Producing quality segmentation masks for images is a fundamental problem in computer vision. Recent research has explored large-scale supervised training to enable zero-shot transfer segmentation on virtually any image style and unsupervised training to enable segmentation without dense annotations. However constructing a model capable of segmenting anything in a zero-shot manner without any annotations is still challenging. In this paper we propose to utilize the self-attention layers in stable diffusion models to achieve this goal because the pre-trained stable diffusion model has learned inherent concepts of objects within its attention layers. Specifically we introduce a simple yet effective iterative merging process based on measuring KL divergence among attention maps to merge them into valid segmentation masks. The proposed method does not require any training or language dependency to extract quality segmentation for any images. On COCO-Stuff-27 our method surpasses the prior unsupervised zero-shot transfer SOTA method by an absolute 26% in pixel accuracy and 17% in mean IoU.
****
TransNeXt: Robust Foveal Visual Perception for Vision Transformers
Dai Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17773-17783
Due to the depth degradation effect in residual connections many efficient Vision Transformers models that rely on stacking layers for information exchange often fail to form sufficient information mixing leading to unnatural visual perception. To address this issue in this paper we propose Aggregated Attention a biomimetic design-based token mixer that simulates biological foveal vision and continuous eye movement while enabling each token on the feature map to have a global perception. Furthermore we incorporate learnable tokens that interact with conventional queries and keys which further diversifies the generation of affinity matrices beyond merely relying on the similarity between queries and keys. Our approach does not rely on stacking for information exchange thus effectively avoiding depth degradation and achieving natural visual perception. Additionally we propose Convolutional GLU a channel mixer that bridges the gap between GLU and SE mechanism which empowers each token to have channel attention based on its nearest neighbor image features enhancing local modeling capability and model robustness. We combine aggregated attention and convolutional GLU to create a new visual backbone called TransNeXt. Extensive experiments demonstrate that our TransNeXt achieves state-of-the-art performance across multiple model sizes. At a resolution of 224^2 TransNeXt-Tiny attains an ImageNet accuracy of 84.0% surpassing C

onvNeXt-B with 69% fewer parameters. Our TransNeXt-Base achieves an ImageNet accuracy of 86.2% and an ImageNet-A accuracy of 61.6% at a resolution of 384^2 a COCO object detection mAP of 57.1 and an ADE20K semantic segmentation mIoU of 54.7.
****
Implicit Discriminative Knowledge Learning for Visible-Infrared Person Re-Identification
Kaijie Ren, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 393-402
Visible-Infrared Person Re-identification (VI-ReID) is a challenging cross-modal pedestrian retrieval task due to significant intra-class variations and cross-modal discrepancies among different cameras. Existing works mainly focus on embedding images of different modalities into a unified space to mine modality-shared features. They only seek distinctive information within these shared features while ignoring the identity-aware useful information that is implicit in the modality-specific features. To address this issue we propose a novel Implicit Discriminative Knowledge Learning (IDKL) network to uncover and leverage the implicit discriminative information contained within the modality-specific. First we extract modality-specific and modality-shared features using a novel dual-stream network. Then the modality-specific features undergo purification to reduce their modality style discrepancies while preserving identity-aware discriminative knowledge. Subsequently this kind of implicit knowledge is distilled into the modality-shared feature to enhance its distinctiveness. Finally an alignment loss is proposed to minimize modality discrepancy on enhanced modality-shared features. Extensive experiments on multiple public datasets demonstrate the superiority of IDKL network over the state-of-the-art methods.
****
Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction
Guillaume Jaume, Anurag Vaidya, Richard J. Chen, Drew F.K. Williamson, Paul Pu Liang, Faisal Mahmood; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11579-11590
Integrating whole-slide images (WSIs) and bulk transcriptomics for predicting patient survival can improve our understanding of patient prognosis. However this multimodal task is particularly challenging due to the different nature of these data: WSIs represent a very high-dimensional spatial description of a tumor while bulk transcriptomics represent a global description of gene expression levels within that tumor. In this context our work aims to address two key challenges: (1) how can we tokenize transcriptomics in a semantically meaningful and interpretable way? and (2) how can we capture dense multimodal interactions between these two modalities? Here we propose to learn biological pathway tokens from transcriptomics that can encode specific cellular functions. Together with histology patch tokens that encode the slide morphology we argue that they form appropriate reasoning units for interpretability. We fuse both modalities using a memory-efficient multimodal Transformer that can model interactions between pathway and histology patch tokens. Our model SURVPATH achieves state-of-the-art performance when evaluated against unimodal and multimodal baselines on five datasets from The Cancer Genome Atlas. Our interpretability framework identifies key multimodal prognostic factors and as such can provide valuable insights into the interaction between genotype and phenotype. Code available at https://github.com/mahmoodlab/SurvPath.
****
Mining Supervision for Dynamic Regions in Self-Supervised Monocular Depth Estimation
Hoang Chuong Nguyen, Tianyu Wang, Jose M. Alvarez, Miaomiao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10446-10455
This paper focuses on self-supervised monocular depth estimation in dynamic scenes trained on monocular videos. Existing methods jointly estimate pixel-wise depth and motion relying mainly on an image reconstruction loss. Dynamic regions re

main a critical challenge for these methods due to the inherent ambiguity in dep th and motion estimation resulting in inaccurate depth estimation. This paper pr oposes a self-supervised training framework exploiting pseudo depth labels for d ynamic regions from training data. The key contribution of our framework is to d ecouple depth estimation for static and dynamic regions of images in the trainin g data. We start with an unsupervised depth estimation approach which provides r eliable depth estimates for static regions and motion cues for dynamic regions a nd allows us to extract moving object information at the instance level. In the next stage we use an object network to estimate the depth of those moving object s assuming rigid motions. Then we propose a new scale alignment module to addres s the scale ambiguity between estimated depths for static and dynamic regions. W e can then use the depth labels generated to train an end-to-end depth estimatio n network and improve its performance. Extensive experiments on the Cityscapes a nd KITTI datasets show that our self-training strategy consistently outperforms existing self-/unsupervised depth estimation methods.
****
Gradient Alignment for Cross-Domain Face Anti-Spoofing
Binh M. Le, Simon S. Woo; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2024, pp. 188-199
Recent advancements in domain generalization (DG) for face anti-spoofing (FAS) h ave garnered considerable attention. Traditional methods have focused on designi ng learning objectives and additional modules to isolate domain-specific feature s while retaining domain-invariant characteristics in their representations. How ever such approaches often lack guarantees of consistent maintenance of domain-i nvariant features or the complete removal of domain-specific features. Furthermo re most prior works of DG for FAS do not ensure convergence to a local flat mini mum which has been shown to be advantageous for DG. In this paper we introduce G AC-FAS a novel learning objective that encourages the model to converge towards an optimal flat minimum without necessitating additional learning modules. Unlik e conventional sharpness-aware minimizers GAC-FAS identifies ascending points fo r each domain and regulates the generalization gradient updates at these points to align coherently with empirical risk minimization (ERM) gradient updates. Thi s unique approach specifically guides the model to be robust against domain shif ts. We demonstrate the efficacy of GAC-FAS through rigorous testing on challengi ng cross-domain FAS datasets where it establishes state-of-the-art performance.
****
Physics-guided Shape-from-Template: Monocular Video Perception through Neural Su rrogate Models
David Stotko, Nils Wandel, Reinhard Klein; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11895-11904
3D reconstruction of dynamic scenes is a long-standing problem in computer graph ics and increasingly difficult the less information is available. Shape-from-Tem plate (SfT) methods aim to reconstruct a template-based geometry from RGB images or video sequences often leveraging just a single monocular camera without dept h information such as regular smartphone recordings. Unfortunately existing reco nstruction methods are either unphysical and noisy or slow in optimization. To s olve this problem we propose a novel SfT reconstruction algorithm for cloth usin g a pre-trained neural surrogate model that is fast to evaluate stable and produ ces smooth reconstructions due to a regularizing physics simulation. Differentia ble rendering of the simulated mesh enables pixel-wise comparisons between the r econstruction and a target video sequence that can be used for a gradient-based optimization procedure to extract not only shape information but also physical p arameters such as stretching shearing or bending stiffness of the cloth. This al lows to retain a precise stable and smooth reconstructed geometry while reducing the runtime by a factor of 400-500 compared to ?-SfT a state-of-the-art physics -based SfT approach.
****
S2MVTC: a Simple yet Efficient Scalable Multi-View Tensor Clustering
Zhen Long, Qiyuan Wang, Yazhou Ren, Yipeng Liu, Ce Zhu; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2621

3-26222
Anchor-based large-scale multi-view clustering has attracted considerable attent
ion for its effectiveness in handling massive datasets. However current methods
mainly seek the consensus embedding feature for clustering by exploring global c
orrelations between anchor graphs or projection matrices.In this paper we propos
e a simple yet efficient scalable multi-view tensor clustering (S2MVTC) approach
 where our focus is on learning correlations of embedding features within and ac
ross views. Specifically we first construct the embedding feature tensor by stac
king the embedding features of different views into a tensor and rotating it. Ad
ditionally we build a novel tensor low-frequency approximation (TLFA) operator w
hich incorporates graph similarity into embedding feature learning efficiently a
chieving smooth representation of embedding features within different views. Fur
thermore consensus constraints are applied to embedding features to ensure inter
-view semantic consistency. Experimental results on six large-scale multi-view d
atasets demonstrate that S2MVTC significantly outperforms state-of-the-art algor
ithms in terms of clustering performance and CPU execution time especially when
handling massive data. The code of S2MVTC is publicly available at https://githu
b.com/longzhen520/S2MVTC.
****
OpticalDR: A Deep Optical Imaging Model for Privacy-Protective Depression Recogn
ition
Yuchen Pan, Junjun Jiang, Kui Jiang, Zhihao Wu, Keyuan Yu, Xianming Liu; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 1303-1312
Depression Recognition (DR) poses a considerable challenge especially in the con
text of the growing concerns surrounding privacy. Traditional automatic diagnosi
s of DR technology necessitates the use of facial images undoubtedly expose the
patient identity features and poses privacy risks. In order to mitigate the pote
ntial risks associated with the inappropriate disclosure of patient facial image
s we design a new imaging system to erase the identity information of captured f
acial images while retain disease-relevant features. It is irreversible for iden
tity information recovery while preserving essential disease-related characteris
tics necessary for accurate DR. More specifically we try to record a de-identifi
ed facial image (erasing the identifiable features as much as possible) by a lea
rnable lens which is optimized in conjunction with the following DR task as well
 as a range of face analysis related auxiliary tasks in an end-to-end manner. Th
ese aforementioned strategies form our final Optical deep Depression Recognition
 network (OpticalDR). Experiments on CelebA AVEC 2013 and AVEC 2014 datasets dem
onstrate that our OpticalDR has achieved state-of-the-art privacy protection per
formance with an average AUC of 0.51 on popular facial recognition models and co
mpetitive results for DR with MAE/RMSE of 7.53/8.48 on AVEC 2013 and 7.89/8.82 o
n AVEC 2014 respectively. Code is available at https://github.com/divertingPan/O
pticalDR.
****
Observation-Guided Diffusion Probabilistic Models
Junoh Kang, Jinyoung Choi, Sungik Choi, Bohyung Han; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8323-83
31
We propose a novel diffusion-based image generation method called the observatio
n-guided diffusion probabilistic model (OGDM) which effectively addresses the tr
adeoff between quality control and fast sampling. Our approach reestablishes the
 training objective by integrating the guidance of the observation process with
the Markov chain in a principled way. This is achieved by introducing an additio
nal loss term derived from the observation based on a conditional discriminator
on noise level which employs a Bernoulli distribution indicating whether its inp
ut lies on the (noisy) real manifold or not. This strategy allows us to optimize
 the more accurate negative log-likelihood induced in the inference stage especi
ally when the number of function evaluations is limited. The proposed training s
cheme is also advantageous even when incorporated only into the fine-tuning proc
ess and it is compatible with various fast inference strategies since our method

yields better denoising networks using the exactly the same inference procedure without incurring extra computational cost. We demonstrate the effectiveness of our training algorithm using diverse inference techniques on strong diffusion model baselines. Our implementation is available at https://github.com/Junoh-Kang/OGDM_edm.
****
You'll Never Walk Alone: A Sketch and Text Duet for Fine-Grained Image Retrieval
Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16509-16519
Two primary input modalities prevail in image retrieval: sketch and text. While text is widely used for inter-category retrieval tasks sketches have been established as the sole preferred modality for fine-grained image retrieval due to their ability to capture intricate visual details. In this paper we question the reliance on sketches alone for fine-grained image retrieval by simultaneously exploring the fine-grained representation capabilities of both sketch and text orchestrating a duet between the two. The end result enables precise retrievals previously unattainable allowing users to pose ever-finer queries and incorporate attributes like colour and contextual cues from text. For this purpose we introduce a novel compositionality framework effectively combining sketches and text using pre-trained CLIP models while eliminating the need for extensive fine-grained textual descriptions. Last but not least our system extends to novel applications in composed image retrieval domain attribute transfer and fine-grained generation providing solutions for various real-world scenarios.
****
Spatial-Aware Regression for Keypoint Localization
Dongkai Wang, Shiliang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 624-633
Regression-based keypoint localization shows advantages of high efficiency and better robustness to quantization errors than heatmap-based methods. However existing regression-based methods discard the spatial location prior in input image with a global pooling leading to inferior accuracy and are limited to single instance localization tasks. We study the regression-based keypoint localization from a new perspective by leveraging the spatial location prior. Instead of regressing on the pooled feature the proposed Spatial-Aware Regression (SAR) maintains the spatial location map and outputs spatial coordinates and confidence score for each grid which are optimized with a unified objective. Benefited by the location prior these spatial-aware outputs can be efficiently optimized resulting in better localization performance. Moreover incorporating spatial prior makes SAR more general and can be applied into various keypoint localization tasks. We test the proposed method in 4 keypoint localization tasks including single/multi-person 2D/3D pose estimation and the whole-body pose estimation. Extensive experiments demonstrate its promising performance e.g. consistently outperforming recent regressions-based methods.
****
S2MAE: A Spatial-Spectral Pretraining Foundation Model for Spectral Remote Sensing Data
Xuyang Li, Danfeng Hong, Jocelyn Chanussot; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24088-24097
In the expansive domain of computer vision a myriad of pre-trained models are at our disposal. However most of these models are designed for natural RGB images and prove inadequate for spectral remote sensing (RS) images. Spectral RS images have two main traits: (1) multiple bands capturing diverse feature information (2) spatial alignment and consistent spectral sequencing within the spatial-spectral dimension. In this paper we introduce Spatial-SpectralMAE (S2MAE) a specialized pre-trained architecture for spectral RS imagery. S2MAE employs a 3D transformer for masked autoencoder modeling integrating learnable spectral-spatial embeddings with a 90% masking ratio. The model efficiently captures local spectral consistency and spatial invariance using compact cube tokens demonstrating versatility to diverse input characteristics. This adaptability facilitates progressi

ve pretraining on extensive spectral datasets. The effectiveness of S2MAE is val idated through continuous pretraining on two sizable datasets totaling over a mi llion training images. The pre-trained model is subsequently applied to three di stinct downstream tasks with in-depth ablation studies conducted to emphasize it s efficacy.
****
EFormer: Enhanced Transformer towards Semantic-Contour Features of Foreground fo r Portraits Matting
Zitao Wang, Qiguang Miao, Yue Xi, Peipei Zhao; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3880-3889
The portrait matting task aims to extract an alpha matte with complete semantics and finely detailed contours. In comparison to CNN-based approaches transformer s with self-attention module have a better capacity to capture long-range depend encies and low-frequency semantic information of a portrait. However recent rese arch shows that the self-attention mechanism struggles with modeling high-freque ncy contour information and capturing fine contour details which can lead to bia s while predicting the portrait's contours. To deal with this issue we propose E Former to enhance the model's attention towards both the low-frequency semantic and high-frequency contour features. For the high-frequency contours our researc h demonstrates that cross-attention module between different resolutions can gui de our model to allocate attention appropriately to these contour regions. Suppo rted by this we can successfully extract the high-frequency detail information a round the portrait's contours which were previously ignored by self-attention. B ased on the cross-attention module we further build a semantic and contour detec tor (SCD) to accurately capture both the low-frequency semantic and high-frequen cy contour features. And we design a contour-edge extraction branch and semantic extraction branch to extract refined high-frequency contour features and comple te low-frequency semantic information respectively. Finally we fuse the two kind s of features and leverage the segmentation head to generate a predicted portrai t matte. Experiments on VideoMatte240K (JPEG SD Format) and Adobe Image Matting (AIM) datasets demonstrate that EFormer outperforms previous portrait matte meth ods.
****
MultiPly: Reconstruction of Multiple People from Monocular Video in the Wild
Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar H illiges, Jie Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 109-118
We present MultiPly a novel framework to reconstruct multiple people in 3D from monocular in-the-wild videos. Reconstructing multiple individuals moving and int eracting naturally from monocular in-the-wild videos poses a challenging task. A ddressing it necessitates precise pixel-level disentanglement of individuals wit hout any prior knowledge about the subjects. Moreover it requires recovering int ricate and complete 3D human shapes from short video sequences intensifying the level of difficulty. To tackle these challenges we first define a layered neural representation for the entire scene composited by individual human and backgrou nd models. We learn the layered neural representation from videos via our layer- wise differentiable volume rendering. This learning process is further enhanced by our hybrid instance segmentation approach which combines the self-supervised 3D segmentation and the promptable 2D segmentation module yielding reliable inst ance segmentation supervision even under close human interaction. A confidence-g uided optimization formulation is introduced to optimize the human poses and sha pe/appearance alternately. We incorporate effective objectives to refine human p oses via photometric information and impose physically plausible constraints on human dynamics leading to temporally consistent 3D reconstructions with high fid elity. The evaluation of our method shows the superiority over prior art on publ icly available datasets and in-the-wild videos.
****
Unsupervised 3D Structure Inference from Category-Specific Image Collections
Weikang Wang, Dongliang Cao, Florian Bernard; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10704-10714

Understanding 3D object structure from image collections of general object categ
ories remains a long-standing challenge in computer vision. Due to the high rele
vance of image keypoints (e.g. for graph matching controlling generative models
scene understanding etc.) in this work we specifically focus on inferring 3D str
ucture in terms of sparse keypoints. Existing 3D keypoint inference approaches r
ely on strong priors such as spatio-temporal consistency multi-view images of th
e same object 3D shape priors (e.g. templates skeleton) or supervisory signals e
.g. in the form of 2D keypoint annotations. In contrast we propose the first uns
upervised 3D keypoint inference approach that can be trained for general object
categories solely from an inhomogeneous image collection (containing different i
nstances of objects from the same category). Our experiments show that our metho
d not only improves upon unsupervised 2D keypoint inference but more importantly
 it also produces reasonable 3D structure for various object categories both qua
litatively and quantitatively.
****
DiG-IN: Diffusion Guidance for Investigating Networks - Uncovering Classifier Di
fferences Neuron Visualisations and Visual Counterfactual Explanations
Maximilian Augustin, Yannic Neuhaus, Matthias Hein; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11093-11
103
While deep learning has led to huge progress in complex image classification tas
ks like ImageNet unexpected failure modes e.g. via spurious features call into q
uestion how reliably these classifiers work in the wild. Furthermore for safety-
critical tasks the black-box nature of their decisions is problematic and explan
ations or at least methods which make decisions plausible are needed urgently. I
n this paper we address these problems by generating images that optimize a clas
sifier-derived objective using a framework for guided image generation. We analy
ze the decisions of image classifiers by visual counterfactual explanations (VCE
s) detection of systematic mistakes by analyzing images where classifiers maxima
lly disagree and visualization of neurons and spurious features. In this way we
validate existing observations e.g. the shape bias of adversarially robust model
s as well as novel failure modes e.g. systematic errors of zero-shot CLIP classi
fiers. Moreover our VCEs outperform previous work while being more versatile.
****
RepViT: Revisiting Mobile CNN From ViT Perspective
Ao Wang, Hui Chen, Zijia Lin, Jungong Han, Guiguang Ding; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15
909-15920
Recently lightweight Vision Transformers (ViTs) demonstrate superior performance
 and lower latency compared with lightweight Convolutional Neural Networks (CNNs
) on resource-constrained mobile devices. Researchers have discovered many struc
tural connections between lightweight ViTs and lightweight CNNs. However the not
able architectural disparities in the block structure macro and micro designs be
tween them have not been adequately examined. In this study we revisit the effic
ient design of lightweight CNNs from ViT perspective and emphasize their promisi
ng prospect for mobile devices. Specifically we incrementally enhance the mobile
-friendliness of a standard lightweight CNN i.e. MobileNetV3 by integrating the
efficient architectural designs of lightweight ViTs. This ends up with a new fam
ily of pure lightweight CNNs namely RepViT. Extensive experiments show that RepV
iT outperforms existing state-of-the-art lightweight ViTs and exhibits favorable
 latency in various vision tasks. Notably on ImageNet RepViT achieves over 80% t
op-1 accuracy with 1.0 ms latency on an iPhone 12 which is the first time for a
lightweight model to the best of our knowledge. Besides when RepViT meets SAM ou
r RepViT-SAM can achieve nearly 10xfaster inference than the advanced MobileSAM.
 Codes and models are available at https://github.com/THU-MIG/RepViT.
****
MonoNPHM: Dynamic Head Reconstruction from Monocular Videos
Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes
Agapito, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2024, pp. 10747-10758

We present Monocular Neural Parametric Head Models (MonoNPHM) for dynamic 3D hea
d reconstructions from monocular RGB videos. To this end we propose a latent app
earance space that parameterizes a texture field on top of a neural parametric m
odel. We constrain predicted color values to be correlated with the underlying g
eometry such that gradients from RGB effectively influence latent geometry codes
 during inverse rendering. To increase the representational capacity of our expr
ession space we augment our backward deformation field with hyper-dimensions thu
s improving color and geometry representation in topologically challenging expre
ssions. Using MonoNPHM as a learned prior we approach the task of 3D head recons
truction using signed distance field based volumetric rendering. By numerically
inverting our backward deformation field we incorporated a landmark loss using f
acial anchor points that are closely tied to our canonical geometry representati
on. We incorporate a facial landmark loss by numerically inverting our backward
deformation field tied with our canonical geometry to observed 2D facial landmar
ks in posed space. To evaluate the task of dynamic face reconstruction from mono
cular RGB videos we record 20 challenging Kinect sequences under casual conditio
ns. MonoNPHM outperforms all baselines with a significant margin and makes an im
portant step towards easily accessible neural parametric face models through RGB
 tracking.
****
Realigning Confidence with Temporal Saliency Information for Point-Level Weakly-
Supervised Temporal Action Localization
Ziying Xia, Jian Cheng, Siyu Liu, Yongxiang Hu, Shiguang Wang, Yijie Zhang, Liwa
n Dang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2024, pp. 18440-18450
Point-level weakly-supervised temporal action localization (P-TAL) aims to local
ize action instances in untrimmed videos through the use of single-point annotat
ions in each instance. Existing methods predict the class activation sequences w
ithout any boundary information and the unreliable sequences result in a signifi
cant misalignment between the quality of proposals and their corresponding confi
dence. In this paper we surprisingly observe the most salient frame tend to appe
ar in the central region of the each instance and is easily annotated by humans.
 Guided by the temporal saliency information we present a novel proposal-level p
lug-in framework to relearn the aligned confidence of proposals generated by the
 base locators. The proposed approach consists of Center Score Learning (CSL) an
d Alignment-based Boundary Adaptation (ABA). In CSL we design a novel center lab
el generated by the point annotations for predicting aligned center scores. Duri
ng inference we first fuse the center scores with the predicted action probabili
ties to obtain the aligned confidence. ABA utilizes the both aligned confidence
and IoU information to enhance localization completeness. Extensive experiments
demonstrate the generalization and effectiveness of the proposed framework showc
asing state-of-the-art or competitive performances across three benchmarks. Our
code is available at https://github.com/zyxia1009/CVPR2024-TSPNet.
****
ConsistNet: Enforcing 3D Consistency for Multi-view Images Diffusion
Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, Hongdong Li; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7
079-7088
Given a single image of a 3D object this paper proposes a novel method (named Co
nsistNet) that can generate multiple images of the same object as if they are ca
ptured from different viewpoints while the 3D (multi-view) consistencies among t
hose multiple generated images are effectively exploited. Central to our method
is a lightweight multi-view consistency block that enables information exchange
across multiple single-view diffusion processes based on the underlying multi-vi
ew geometry principles. ConsistNet is an extension to the standard latent diffus
ion model and it consists of two submodules: (a) a view aggregation module that
unprojects multi-view features into global 3D volumes and infers consistency and
 (b) a ray aggregation module that samples and aggregates 3D consistent features
 back to each view to enforce consistency. Our approach departs from previous me
thods in multi-view image generation in that it can be easily dropped in pre-tra

ined LDMs without requiring explicit pixel correspondences or depth prediction. Experiments show that our method effectively learns 3D consistency over a frozen Zero123-XL backbone and can generate 16 surrounding views of the object within 11 seconds on a single A100 GPU.
****
GenN2N: Generative NeRF2NeRF Translation
Xiangyue Liu, Han Xue, Kunming Luo, Ping Tan, Li Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5105-5114
We present GenN2N a unified NeRF-to-NeRF translation framework for various NeRF translation tasks such as text-driven NeRF editing colorization super-resolution inpainting etc. Unlike previous methods designed for individual translation tasks with task-specific schemes GenN2N achieves all these NeRF editing tasks by employing a plug-and-play image-to-image translator to perform editing in the 2D domain and lifting 2D edits into the 3D NeRF space. Since the 3D consistency of 2D edits may not be assured we propose to model the distribution of the underlying 3D edits through a generative model that can cover all possible edited NeRFs. To model the distribution of 3D edited NeRFs from 2D edited images we carefully design a VAE-GAN that encodes images while decoding NeRFs. The latent space is trained to align with a Gaussian distribution and the NeRFs are supervised through an adversarial loss on its renderings. To ensure the latent code does not depend on 2D viewpoints but truly reflects the 3D edits we also regularize the latent code through a contrastive learning scheme. Extensive experiments on various editing tasks show GenN2N as a universal framework performs as well or better than task-specific specialists while possessing flexible generative power. More results on our project page: https://xiangyueliu.github.io/GenN2N/.
****
Theoretically Achieving Continuous Representation of Oriented Bounding Boxes
Zikai Xiao, Guoye Yang, Xue Yang, Taijiang Mu, Junchi Yan, Shimin Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16912-16922
Considerable efforts have been devoted to Oriented Object Detection (OOD). However one lasting issue regarding the discontinuity in Oriented Bounding Box (OBB) representation remains unresolved which is an inherent bottleneck for extant OOD methods. This paper endeavors to completely solve this issue in a theoretically guaranteed manner and puts an end to the ad-hoc efforts in this direction. Prior studies typically can only address one of the two cases of discontinuity: rotation and aspect ratio and often inadvertently introduce decoding discontinuity e.g. Decoding Incompleteness (DI) and Decoding Ambiguity (DA) as discussed in literature. Specifically we propose a novel representation method called Continuous OBB (COBB) which can be readily integrated into existing detectors e.g. Faster-RCNN as a plugin. It can theoretically ensure continuity in bounding box regression which to our best knowledge has not been achieved in literature for rectangle-based object representation. For fairness and transparency of experiments we have developed a modularized benchmark based on the open-source deep learning framework Jittor's detection toolbox JDet for OOD evaluation. On the popular DOTA dataset by integrating Faster-RCNN as the same baseline model our new method outperforms the peer method Gliding Vertex by 1.13% mAP50 (relative improvement 1.54%) and 2.46% mAP75 (relative improvement 5.91%) without any tricks.
****
Universal Robustness via Median Randomized Smoothing for Real-World Super-Resolution
Zakariya Chaouai, Mohamed Tamaazousti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9059-9068
Most of the recent literature on image Super-Resolution (SR) can be classified into two main approaches. The first one involves learning a corruption model tailored to a specific dataset aiming to mimic the noise and corruption in low-resolution images such as sensor noise. However this approach is data-specific tends to lack adaptability and its accuracy diminishes when faced with unseen types of image corruptions. A second and more recent approach referred to as Robust Supe

r-Resolution (RSR) proposes to improve real-world SR by harnessing the generaliz ation capabilities of a model by making it robust to adversarial attacks. To del ve further into this second approach our paper explores the universality of vari ous methods for enhancing the robustness of deep learning SR models. In other wo rds we inquire: \enquote Which robustness method exhibits the highest degree of adaptability when dealing with a wide range of adversarial attacks ? . Our exten sive experimentation on both synthetic and real-world images empirically demonst rates that median randomized smoothing (MRS) is more general in terms of robustn ess compared to adversarial learning techniques which tend to focus on specific types of attacks. Furthermore as expected we also illustrate that the proposed u niversal robust method enables the SR model to handle standard corruptions more effectively such as blur and Gaussian noise and notably corruptions naturally pr esent in real-world images. These results support the significance of shifting t he paradigm in the development of real-world SR methods towards RSR especially v ia MRS.
****

One-dimensional Adapter to Rule Them All: Concepts Diffusion Models and Erasing Applications

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jun gong Han, Guiguang Ding; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2024, pp. 7559-7568

The prevalent use of commercial and open-source diffusion models (DMs) for text-to-image generation prompts risk mitigation to prevent undesired behaviors. Exis ting concept erasing methods in academia are all based on full parameter or spec ification-based fine-tuning from which we observe the following issues: 1) Gener ation alteration towards erosion: Parameter drift during target elimination caus es alterations and potential deformations across all generations even eroding ot her concepts at varying degrees which is more evident with multi-concept erased; 2) Transfer inability & deployment inefficiency: Previous model-specific erasur e impedes the flexible combination of concepts and the training-free transfer to wards other models resulting in linear cost growth as the deployment scenarios i ncrease. To achieve non-invasive precise customizable and transferable eliminati on we ground our erasing framework on one-dimensional adapters to erase multiple concepts from most DMs at once across versatile erasing applications. The conce pt-SemiPermeable structure is injected as a Membrane (SPM) into any DM to learn targeted erasing and meantime the alteration and erosion phenomenon is effective ly mitigated via a novel Latent Anchoring fine-tuning strategy. Once obtained SP Ms can be flexibly combined and plug-and-play for other DMs without specific re-tuning enabling timely and efficient adaptation to diverse scenarios. During gen eration our Facilitated Transport mechanism dynamically regulates the permeabili ty of each SPM to respond to different input prompts further minimizing the impa ct on other concepts. Quantitative and qualitative results across 40 concepts 7 DMs and 4 erasing applications have demonstrated the superior erasing of SPM. O ur code and pre-tuned SPMs are available on the project page https://lyumengyao. github.io/projects/spm.
****

Learning Large-Factor EM Image Super-Resolution with Generative Priors

Jiateng Shou, Zeyu Xiao, Shiyu Deng, Wei Huang, Peiyao Shi, Ruobing Zhang, Zhiwe i Xiong, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11313-11322

As the mainstream technique for capturing images of biological specimens at nano meter resolution electron microscopy (EM) is extremely time-consuming for scanni ng wide field-of-view (FOV) specimens. In this paper we investigate a challengin g task of large-factor EM image super-resolution (EMSR) which holds great promis e for reducing scanning time relaxing acquisition conditions and expanding imagi ng FOV. By exploiting the repetitive structures and volumetric coherence of EM i mages we propose the first generative learning-based framework for large-factor EMSR. Specifically motivated by the predictability of repetitive structures and textures in EM images we first learn a discrete codebook in the latent space to represent high-resolution (HR) cell-specific priors and a latent vector indexer

to map low-resolution (LR) EM images to their corresponding latent vectors in a generative manner. By incorporating the generative cell-specific priors from HR EM images through a multi-scale prior fusion module we then deploy multi-image feature alignment and fusion to further exploit the inter-section coherence in the volumetric EM data. Extensive experiments demonstrate that our proposed framework outperforms advanced single-image and video super-resolution methods for 8x and 16x EMSR (i.e. with 64 times and 256 times less data acquired respectively) achieving superior visual reconstruction quality and downstream segmentation accuracy on benchmark EM datasets. Code is available at https://github.com/jtshou/GPEMSR.
****
DIMAT: Decentralized Iterative Merging-And-Training for Deep Learning Models
Nastaran Saadati, Minh Pham, Nasla Saleem, Joshua R. Waite, Aditya Balu, Zhanong Jiang, Chinmay Hegde, Soumik Sarkar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27517-27527
Recent advances in decentralized deep learning algorithms have demonstrated cutting-edge performance on various tasks with large pre-trained models. However a pivotal prerequisite for achieving this level of competitiveness is the significant communication and computation overheads when updating these models which prohibits the applications of them to real-world scenarios. To address this issue drawing inspiration from advanced model merging techniques without requiring additional training we introduce the Decentralized Iterative Merging-And-Training (DIMAT) paradigm--a novel decentralized deep learning framework. Within DIMAT each agent is trained on their local data and periodically merged with their neighboring agents using advanced model merging techniques like activation matching until convergence is achieved. DIMAT provably converges with the best available rate for nonconvex functions with various first-order methods while yielding tighter error bounds compared to the popular existing approaches. We conduct a comprehensive empirical analysis to validate DIMAT's superiority over baselines across diverse computer vision tasks sourced from multiple datasets. Empirical results validate our theoretical claims by showing that DIMAT attains faster and higher initial gain in accuracy with independent and identically distributed (IID) and non-IID data incurring lower communication overhead. This DIMAT paradigm presents a new opportunity for the future decentralized learning enhancing its adaptability to real-world with sparse and light-weight communication and computation.
****
MMA: Multi-Modal Adapter for Vision-Language Models
Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, Xiaohua Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23826-23837
Pre-trained Vision-Language Models (VLMs) have served as excellent foundation models for transfer learning in diverse downstream tasks. However tuning VLMs for few-shot generalization tasks faces a discrimination -- generalization dilemma i.e. general knowledge should be preserved and task-specific knowledge should be fine-tuned. How to precisely identify these two types of representations remains a challenge. In this paper we propose a Multi-Modal Adapter (MMA) for VLMs to improve the alignment between representations from text and vision branches. MMA aggregates features from different branches into a shared feature space so that gradients can be communicated across branches. To determine how to incorporate MMA we systematically analyze the discriminability and generalizability of features across diverse datasets in both the vision and language branches and find that (1) higher layers contain discriminable dataset-specific knowledge while lower layers contain more generalizable knowledge and (2) language features are more discriminable than visual features and there are large semantic gaps between the features of the two modalities especially in the lower layers. Therefore we only incorporate MMA to a few higher layers of transformers to achieve an optimal balance between discrimination and generalization. We evaluate the effectiveness of our approach on three tasks: generalization to novel classes novel target datasets and domain generalization. Compared to many state-of-the-art methods our MMA achieves leading performance in all evaluations. Code is at https://github.co

m/ZjjConan/Multi-Modal-Adapter
****
Kandinsky Conformal Prediction: Efficient Calibration of Image Segmentation Algorithms
Joren Brunekreef, Eric Marcus, Ray Sheombarsing, Jan-Jakob Sonke, Jonas Teuwen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4135-4143
Image segmentation algorithms can be understood as a collection of pixel classifiers for which the outcomes of nearby pixels are correlated. Classifier models can be calibrated using Inductive Conformal Prediction but this requires holding back a sufficiently large calibration dataset for computing the distribution of non-conformity scores of the model's predictions. If one only requires only marginal calibration on the image level this calibration set consists of all individual pixels in the images available for calibration. However if the goal is to attain proper calibration for each individual pixel classifier the calibration set consists of individual images. In a scenario where data are scarce (such as the medical domain) it may not always be possible to set aside sufficiently many images for this pixel-level calibration.The method we propose dubbed "Kandinsky calibration" makes use of the spatial structure present in the distribution of natural images to simultaneously calibrate the classifiers of "similar" pixels. This can be seen as an intermediate approach between marginal (imagewise) and conditional (pixelwise) calibration where non-conformity scores are aggregated over similar image regions thereby making more efficient use of the images available for calibration. We run experiments on segmentation algorithms trained and calibrated on subsets of the public MS-COCO and Medical Decathlon datasets demonstrating that Kandinsky calibration method can significantly improve the coverage. When compared to both pixelwise and imagewise calibration on little data the Kandinsky method achieves much lower coverage errors indicating the data efficiency of the Kandinsky calibration.
****
Diversity-aware Channel Pruning for StyleGAN Compression
Jiwoo Chung, Sangeek Hyun, Sang-Heon Shim, Jae-Pil Heo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7902-7911
StyleGAN has shown remarkable performance in unconditional image generation. However its high computational cost poses a significant challenge for practical applications. Although recent efforts have been made to compress StyleGAN while preserving its performance existing compressed models still lag behind the original model particularly in terms of sample diversity. To overcome this we propose a novel channel pruning method that leverages varying sensitivities of channels to latent vectors which is a key factor in sample diversity. Specifically by assessing channel importance based on their sensitivities to latent vector perturbations our method enhances the diversity of samples in the compressed model. Since our method solely focuses on the channel pruning stage it has complementary benefits with prior training schemes without additional training cost. Extensive experiments demonstrate that our method significantly enhances sample diversity across various datasets. Moreover in terms of FID scores our method not only surpasses state-of-the-art by a large margin but also achieves comparable scores with only half training iterations.
****
BioCLIP: A Vision Foundation Model for the Tree of Life
Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, Yu Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19412-19424
Images of the natural world collected by a variety of cameras from drones to individual phones are increasingly abundant sources of biological information. There is an explosion of computational methods and tools particularly computer vision for extracting biologically relevant information from images for science and conservation. Yet most of these are bespoke approaches designed for a specific ta

sk and are not easily adaptable or extendable to new questions contexts and data sets. A vision model for general organismal biology questions on images is of timely need. To approach this we curate and release TreeOfLife-10M the largest and most diverse ML-ready dataset of biology images. We then develop BioCLIP a foundation model for the tree of life leveraging the unique properties of biology captured by TreeOfLife-10M namely the abundance and variety of images of plants animals and fungi together with the availability of rich structured biological knowledge. We rigorously benchmark our approach on diverse fine-grained biology classification tasks and find that BioCLIP consistently and substantially outperforms existing baselines (by 17% to 20% absolute). Intrinsic evaluation reveals that BioCLIP has learned a hierarchical representation conforming to the tree of life shedding light on its strong generalizability. All data code and models will be publicly released upon acceptance.
****
From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models
Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, Xuming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28076-28086
Scene graph generation (SGG) aims to parse a visual scene into an intermediate graph representation for downstream reasoning tasks. Despite recent advancements existing methods struggle to generate scene graphs with novel visual relation concepts. To address this challenge we introduce a new open-vocabulary SGG framework based on sequence generation. Our framework leverages vision-language pre-trained models (VLM) by incorporating an image-to-graph generation paradigm. Specifically we generate scene graph sequences via image-to-text generation with VLM and then construct scene graphs from these sequences. By doing so we harness the strong capabilities of VLM for open-vocabulary SGG and seamlessly integrate explicit relational modeling for enhancing the VL tasks. Experimental results demonstrate that our design not only achieves superior performance with an open vocabulary but also enhances downstream vision-language task performance through explicit relation modeling knowledge.
****
Deep Imbalanced Regression via Hierarchical Classification Adjustment
Haipeng Xiong, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23721-23730
Regression tasks in computer vision such as age estimation or counting are often formulated into classification by quantizing the target space into classes. Yet real-world data is often imbalanced -- the majority of training samples lie in a head range of target values while a minority of samples span a usually larger tail range. By selecting the class quantization one can adjust imbalanced regression targets into balanced classification outputs though there are trade-offs in balancing classification accuracy and quantization error. To improve regression performance over the entire range of data we propose to construct hierarchical classifiers for solving imbalanced regression tasks. The fine-grained classifiers limit the quantization error while being modulated by the coarse predictions to ensure high accuracy. Standard hierarchical classification approaches when applied to the regression problem fail to ensure that predicted ranges remain consistent across the hierarchy. As such we propose a range-preserving distillation process that effectively learns a single classifier from the set of hierarchical classifiers. Our novel hierarchical classification adjustment (HCA) for imbalanced regression shows superior results on three diverse tasks: age estimation crowd counting and depth estimation. Code is available at https://github.com/xhp-hust-2018-2011/HCA.
****
Adaptive Fusion of Single-View and Multi-View Depth for Autonomous Driving
Junda Cheng, Wei Yin, Kaixuan Wang, Xiaozhi Chen, Shijie Wang, Xin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10138-10147
Multi-view depth estimation has achieved impressive performance over various ben

chmarks. However almost all current multi-view systems rely on given ideal camer a poses which are unavailable in many real-world scenarios such as autonomous dr iving. In this work we propose a new robustness benchmark to evaluate the depth estimation system under various noisy pose settings. Surprisingly we find curren t multi-view depth estimation methods or single-view and multi-view fusion metho ds will fail when given noisy pose settings. To address this challenge we propos e a single-view and multi-view fused depth estimation system which adaptively in tegrates high-confident multi-view and single-view results for both robust and a ccurate depth estimations. The adaptive fusion module performs fusion by dynamic ally selecting high-confidence regions between two branches based on a wrapping confidence map. Thus the system tends to choose the more reliable branch when fa cing textureless scenes inaccurate calibration dynamic objects and other degrada tion or challenging conditions. Our method outperforms state-of-the-art multi-vi ew and fusion methods under robustness testing. Furthermore we achieve state-of- the-art performance on challenging benchmarks (KITTI and DDAD) when given accura te pose estimations. Project website: https://github.com/Junda24/AFNet/
****
Neural Clustering based Visual Representation Learning
Guikun Chen, Xia Li, Yi Yang, Wenguan Wang; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5714-5725
We investigate a fundamental aspect of machine vision: the measurement of featur es by revisiting clustering one of the most classic approaches in machine learni ng and data analysis. Existing visual feature extractors including ConvNets ViTs and MLPs represent an image as rectangular regions. Though prevalent such a gri d-style paradigm is built upon engineering practice and lacks explicit modeling of data distribution. In this work we propose feature extraction with clustering (FEC) a conceptually elegant yet surprisingly ad-hoc interpretable neural clust ering framework which views feature extraction as a process of selecting represe ntatives from data and thus automatically captures the underlying data distribut ion. Given an image FEC alternates between grouping pixels into individual clust ers to abstract representatives and updating the deep features of pixels with cu rrent representatives. Such an iterative working mechanism is implemented in the form of several neural layers and the final representatives can be used for dow nstream tasks. The cluster assignments across layers which can be viewed and ins pected by humans make the forward process of FEC fully transparent and empower i t with promising ad-hoc interpretability. Extensive experiments on various visua l recognition models and tasks verify the effectiveness generality and interpret ability of FEC. We expect this work will provoke a rethink of the current de fac to grid-style paradigm.
****
Continual Self-supervised Learning: Towards Universal Multi-modal Medical Data R epresentation Learning
Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Qi Wu, Yong Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 24, pp. 11114-11124
Self-supervised learning (SSL) is an efficient pre-training method for medical i mage analysis. However current research is mostly confined to certain modalities consuming considerable time and resources without achieving universality across different modalities. A straightforward solution is combining all modality data for joint SSL which poses practical challenges. Firstly our experiments reveal conflicts in representation learning as the number of modalities increases. Seco ndly multi-modal data collected in advance cannot cover all real-world scenarios . In this paper we reconsider versatile SSL from the perspective of continual le arning and propose MedCoSS a continuous SSL approach for multi-modal medical dat a. Different from joint representation learning MedCoSS assigns varying data mod alities to separate training stages creating a multi-stage pre-training process. We propose a rehearsal-based continual learning approach to manage modal confli cts and prevent catastrophic forgetting. Specifically we use the k-means samplin g to retain and rehearse previous modality data during new modality learning. Mo reover we apply feature distillation and intra-modal mixup on buffer data for kn

owledge retention bypassing pretext tasks. We conduct experiments on a large-sca
le multi-modal unlabeled dataset including clinical reports X-rays CT MRI and pa
thological images. Experimental results demonstrate MedCoSS's exceptional genera
lization ability across 9 downstream datasets and its significant scalability in
 integrating new modality data. The code and pre-trained model are available at
https://github.com/yeerwen/MedCoSS.
****
Sparse Semi-DETR: Sparse Learnable Queries for Semi-Supervised Object Detection
Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, Muhammad Zeshan Afzal; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 5840-5850
In this paper we address the limitations of the DETR-based semi-supervised objec
t detection (SSOD) framework particularly focusing on the challenges posed by th
e quality of object queries. In DETR-based SSOD the one-to-one assignment strate
gy provides inaccurate pseudo-labels while the one-to-many assignments strategy
leads to overlapping predictions. These issues compromise training efficiency an
d degrade model performance especially in detecting small or occluded objects. W
e introduce Sparse Semi-DETR a novel transformer-based end-to-end semi-supervise
d object detection solution to overcome these challenges. Sparse Semi-DETR incor
porates a Query Refinement Module to enhance the quality of object queries signi
ficantly improving detection capabilities for small and partially obscured objec
ts. Additionally we integrate a Reliable Pseudo-Label Filtering Module that sele
ctively filters high-quality pseudo-labels thereby enhancing detection accuracy
and consistency. On the MS-COCO and Pascal VOC object detection benchmarks Spars
e Semi-DETR achieves a significant improvement over current state-of-the-art met
hods that highlight Sparse Semi-DETR's effectiveness in semi-supervised object d
etection particularly in challenging scenarios involving small or partially obsc
ured objects.
****
Towards Efficient Replay in Federated Incremental Learning
Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, Guannan Zhang; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2024, pp. 12820-12829
In Federated Learning (FL) the data in each client is typically assumed fixed or
 static. However data often comes in an incremental manner in real-world applica
tions where the data domain may increase dynamically. In this work we study cata
strophic forgetting with data heterogeneity in Federated Incremental Learning (F
IL) scenarios where edge clients may lack enough storage space to retain full da
ta. We propose to employ a simple generic framework for FIL named Re-Fed which c
an coordinate each client to cache important samples for replay. More specifical
ly when a new task arrives each client first caches selected previous samples ba
sed on their global and local importance. Then the client trains the local model
 with both the cached samples and the samples from the new task. Theoretically w
e analyze the ability of Re-Fed to discover important samples for replay thus al
leviating the catastrophic forgetting problem. Moreover we empirically show that
 Re-Fed achieves competitive performance compared to state-of-the-art methods.
****
SimAC: A Simple Anti-Customization Method for Protecting Face Privacy against Te
xt-to-Image Synthesis of Diffusion Models
Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, Qidong Huang; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
12047-12056
Despite the success of diffusion-based customization methods on visual content c
reation increasing concerns have been raised about such techniques from both pri
vacy and political perspectives. To tackle this issue several anti-customization
 methods have been proposed in very recent months predominantly grounded in adve
rsarial attacks. Unfortunately most of these methods adopt straightforward desig
ns such as end-to-end optimization with a focus on adversarially maximizing the
original training loss thereby neglecting nuanced internal properties intrinsic
to the diffusion model and even leading to ineffective optimization in some diff

usion time steps. In this paper we strive to bridge this gap by undertaking a comprehensive exploration of these inherent properties to boost the performance of current anti-customization approaches. Two aspects of properties are investigated: 1) We examine the relationship between time step selection and the model's perception in the frequency domain of images and find that lower time steps can give much more contributions to adversarial noises. This inspires us to propose an adaptive greedy search for optimal time steps that seamlessly integrates with existing anti-customization methods. 2) We scrutinize the roles of features at different layers during denoising and devise a sophisticated feature-based optimization framework for anti-customization. Experiments on facial benchmarks demonstrate that our approach significantly increases identity disruption thereby protecting user privacy and copyright. Our code is available at: https://github.com/somuchtome/SimAC.
****
Total-Decom: Decomposed 3D Scene Reconstruction with Minimal Interaction
Xiaoyang Lyu, Chirui Chang, Peng Dai, Yang-Tian Sun, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20860-20869
Scene reconstruction from multi-view images is a fundamental problem in computer vision and graphics. Recent neural implicit surface reconstruction methods have achieved high-quality results; however editing and manipulating the 3D geometry of reconstructed scenes remains challenging due to the absence of naturally decomposed object entities and complex object/background compositions. In this paper we present Total-Decom a novel method for decomposed 3D reconstruction with minimal human interaction. Our approach seamlessly integrates the Segment Anything Model (SAM) with hybrid implicit-explicit neural surface representations and a mesh-based region-growing technique for accurate 3D object decomposition. Total-Decom requires minimal human annotations while providing users with real-time control over the granularity and quality of decomposition. We extensively evaluate our method on benchmark datasets and demonstrate its potential for downstream applications such as animation and scene editing.
****
Accelerating Neural Field Training via Soft Mining
Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, Kwang Moo Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20071-20080
We present an approach to accelerate Neural Field training by efficiently selecting sampling locations. While Neural Fields have recently become popular it is often trained by uniformly sampling the training domain or through handcrafted heuristics. We show that improved convergence and final training quality can be achieved by a soft mining technique based on importance sampling: rather than either considering or ignoring a pixel completely we weigh the corresponding loss by a scalar. To implement our idea we use Langevin Monte-Carlo sampling. We show that by doing so regions with higher error are being selected more frequently leading to more than 2x improvement in convergence speed. The code and related resources for this study are publicly available at https://ubc-vision.github.io/nf-soft-mining/.
****
Ensemble Diversity Facilitates Adversarial Transferability
Bowen Tang, Zheng Wang, Yi Bin, Qi Dou, Yang Yang, Heng Tao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24377-24386
With the advent of ensemble-based attacks the transferability of generated adversarial examples is elevated by a noticeable margin despite many methods only employing superficial integration yet ignoring the diversity between ensemble models. However most of them compromise the latent value of the diversity between generated perturbation from distinct models which we argue is also able to increase the adversarial transferability especially heterogeneous attacks. To address the issues we propose a novel method of Stochastic Mini-batch black-box attack with Ensemble Reweighing using reinforcement learning (SMER) to produce highly tran

sferable adversarial examples. We emphasize the diversity between surrogate models achieving individual perturbation iteratively. In order to customize the individual effect between surrogates ensemble reweighing is introduced to refine ensemble weights by maximizing attack loss based on reinforcement learning which functions on the ultimate transferability elevation. Extensive experiments demonstrate our superiority to recent ensemble attacks with a significant margin across different black-box attack scenarios especially on heterogeneous conditions.
****

Fair-VPT: Fair Visual Prompt Tuning for Image Classification
Sungho Park, Hyeran Byun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12268-12278
Despite the remarkable success of Vision Transformers (ViT) across diverse fields in computer vision they have a clear drawback of expensive adaption cost for downstream tasks due to the increased scale. To address this Visual Prompt Tuning (VPT) incorporates learnable parameters in the input space of ViT. While freezing the ViT backbone and tuning only the prompts it exhibits superior performances to full fine-tuning. However despite the outstanding advantage we point out that VPT may lead to serious unfairness in downstream classification. Initially we investigate the causes of unfairness in VPT identifying the biasedly pre-trained ViT as a principal factor. Motivated by this observation we propose a Fair Visual Prompt Tuning (Fair-VPT) which removes biased information in the pre-trained ViT while adapting it to downstream classification tasks. To this end we categorize prompts into "cleaner prompts" and "target prompts". Based on this we encode the class token in two different ways by either masking or not masking the target prompts in the self-attention process. These encoded tokens are trained with distinct objective functions resulting in the inclusion of different information in the target and cleaner prompts. Moreover we introduce a disentanglement loss based on contrastive learning to further decorrelate them. In experiments across diverse benchmarks the proposed method demonstrates the most superior performance in terms of balanced classification accuracy and fairness.
****

Uncertainty-Aware Source-Free Adaptive Image Super-Resolution with Wavelet Augmentation Transformer
Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Lei Zhang, Ran He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8142-8152
Unsupervised Domain Adaptation (UDA) can effectively address domain gap issues in real-world image Super-Resolution (SR) by accessing both the source and target data. Considering privacy policies or transmission restrictions of source data in practical scenarios we propose a SOurce-free Domain Adaptation framework for image SR (SODA-SR) to address this issue i.e. adapt a source-trained model to a target domain with only unlabeled target data. SODA-SR leverages the source-trained model to generate refined pseudo-labels for teacher-student learning. To better utilize pseudo-labels we propose a novel wavelet-based augmentation method named Wavelet Augmentation Transformer (WAT) which can be flexibly incorporated with existing networks to implicitly produce useful augmented data. WAT learns low-frequency information of varying levels across diverse samples which is aggregated efficiently via deformable attention. Furthermore an uncertainty-aware self-training mechanism is proposed to improve the accuracy of pseudo-labels with inaccurate predictions being rectified by uncertainty estimation. To acquire better SR results and avoid overfitting pseudo-labels several regularization losses are proposed to constrain target LR and SR images in the frequency domain. Experiments show that without accessing source data SODA-SR outperforms state-of-the-art UDA methods in both synthetic->real and real->real adaptation settings and is not constrained by specific network architectures.
****

Gear-NeRF: Free-Viewpoint Rendering and Tracking with Motion-aware Spatio-Temporal Sampling
Xinhang Liu, Yu-Wing Tai, Chi-Keung Tang, Pedro Miraldo, Suhas Lohit, Moitreya Chatterjee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), 2024, pp. 19667-19679
Extensions of Neural Radiance Fields (NeRFs) to model dynamic scenes have enable
d their near photo-realistic free-viewpoint rendering. Although these methods ha
ve shown some potential in creating immersive experiences two drawbacks limit th
eir ubiquity: (i) a significant reduction in reconstruction quality when the com
puting budget is limited and (ii) a lack of semantic understanding of the underl
ying scenes. To address these issues we introduce Gear-NeRF which leverages sema
ntic information from powerful image segmentation models. Our approach presents
a principled way for learning a spatio-temporal (4D) semantic embedding based on
 which we introduce the concept of gears to allow for stratified modeling of dyn
amic regions of the scene based on the extent of their motion. Such differentiat
ion allows us to adjust the spatio-temporal sampling resolution for each region
in proportion to its motion scale achieving more photo-realistic dynamic novel v
iew synthesis. At the same time almost for free our approach enables free-viewpo
int tracking of objects of interest -- a functionality not yet achieved by exist
ing NeRF-based methods. Empirical studies validate the effectiveness of our meth
od where we achieve state-of-the-art rendering and tracking performance on multi
ple challenging datasets. The project page is available at: https://merl.com/res
earch/highlights/gear-nerf.
****
CaDeT: a Causal Disentanglement Approach for Robust Trajectory Prediction in Aut
onomous Driving
Mozhgan Pourkeshavarz, Junrui Zhang, Amir Rasouli; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14874-148
84
For safe motion planning in real-world autonomous vehicles require behavior pred
iction models that are reliable and robust to distribution shifts. The recent st
udies suggest that the existing learning-based trajectory prediction models do n
ot posses such characteristics and are susceptible to small perturbations that a
re not present in the training data largely due to overfitting to spurious corre
lations while learning. In this paper we propose a causal disentanglement repres
entation learning approach aiming to separate invariant (causal) and variant (sp
urious) features for more robust learning. Our method benefits from a novel inte
rvention mechanism in the latent space that estimates potential distribution shi
fts resulted from spurious correlations using uncertain feature statistics hence
 maintaining the realism of interventions. To facilitate learning we propose a n
ovel invariance objective based on the variances of the distributions over uncer
tain statistics to induce the model to focus on invariant representations during
 training. We conduct extensive experiments on two large-scale autonomous drivin
g datasets and show that besides achieving state-of-the-art performance our meth
od can significantly improve prediction robustness to various distribution shift
s in driving scenes. We further conduct ablative studies to evaluate the design
choices in our proposed framework.
****
Spacetime Gaussian Feature Splatting for Real-Time Dynamic View Synthesis
Zhan Li, Zhang Chen, Zhong Li, Yi Xu; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8508-8520
Novel view synthesis of dynamic scenes has been an intriguing yet challenging pr
oblem. Despite recent advancements simultaneously achieving high-resolution phot
orealistic results real-time rendering and compact storage remains a formidable
task. To address these challenges we propose Spacetime Gaussian Feature Splattin
g as a novel dynamic scene representation composed of three pivotal components.
First we formulate expressive Spacetime Gaussians by enhancing 3D Gaussians with
 temporal opacity and parametric motion/rotation. This enables Spacetime Gaussia
ns to capture static dynamic as well as transient content within a scene. Second
 we introduce splatted feature rendering which replaces spherical harmonics with
 neural features. These features facilitate the modeling of view- and time-depen
dent appearance while maintaining small size. Third we leverage the guidance of
training error and coarse depth to sample new Gaussians in areas that are challe
nging to converge with existing pipelines. Experiments on several established re

al-world datasets demonstrate that our method achieves state-of-the-art renderin
g quality and speed while retaining compact storage. At 8K resolution our lite-v
ersion model can render at 60 FPS on an Nvidia RTX 4090 GPU.
****
Instruct-Imagen: Image Generation with Multi-modal Instruction
Hexiang Hu, Kelvin C.K. Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, Ming-Wei Chang, Xuhui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4754-4763
This paper presents Instruct-Imagen a model that tackles heterogeneous image generation tasks and generalizes across unseen tasks. We introduce multi-modal instruction for image generation a task representation articulating a range of generation intents with precision. It uses natural language to amalgamate disparate modalities (e.g. text edge style subject etc.) such that abundant generation intents can be standardized in a uniform format. We then build Instruct-Imagen by fine-tuning a pre-trained text-to-image diffusion model with two stages. First we adapt the model using the retrieval-augmented training to enhance model's capabilities to ground its generation on external multi-modal context. Subsequently we fine-tune the adapted model on diverse image generation tasks that requires vision-language understanding (e.g. subject-driven generation etc.) each paired with a multi-modal instruction encapsulating the task's essence. Human evaluation on various image generation datasets reveals that Instruct-Imagen matches or surpasses prior task-specific models in-domain and demonstrates promising generalization to unseen and more complex tasks. Our evaluation suite will be made publicly available.
****
Prompting Vision Foundation Models for Pathology Image Analysis
Chong Yin, Siqi Liu, Kaiyang Zhou, Vincent Wai-Sun Wong, Pong C. Yuen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11292-11301
The rapid increase in cases of non-alcoholic fatty liver disease (NAFLD) in recent years has raised significant public concern. Accurately identifying tissue alteration regions is crucial for the diagnosis of NAFLD but this task presents challenges in pathology image analysis particularly with small-scale datasets. Recently the paradigm shift from full fine-tuning to prompting in adapting vision foundation models has offered a new perspective for small-scale data analysis. However existing prompting methods based on task-agnostic prompts are mainly developed for generic image recognition which fall short in providing instructive cues for complex pathology images. In this paper we propose Q uantitative A ttribute-based P rompting (QAP) a novel prompting method specifically for liver pathology image analysis. QAP is based on two quantitative attributes namely K-function-based spatial attributes and histogram-based morphological attributes which are aimed for quantitative assessment of tissue states. Moreover a conditional prompt generator is designed to turn these instance-specific attributes into visual prompts. Extensive experiments on three diverse tasks demonstrate that our task-specific prompting method achieves better diagnostic performance as well as better interpretability. Code is available at \href https://github.com/7LFB/QAP https://github.com/7LFB/QAP .
****
Rethinking Few-shot 3D Point Cloud Semantic Segmentation
Zhaochong An, Guolei Sun, Yun Liu, Fayao Liu, Zongwei Wu, Dan Wang, Luc Van Gool, Serge Belongie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3996-4006
This paper revisits few-shot 3D point cloud semantic segmentation (FS-PCS) with a focus on two significant issues in the state-of-the-art: foreground leakage and sparse point distribution. The former arises from non-uniform point sampling allowing models to distinguish the density disparities between foreground and background for easier segmentation. The latter results from sampling only 2048 points limiting semantic information and deviating from the real-world practice. To address these issues we introduce a standardized FS-PCS setting upon which a new

benchmark is built. Moreover we propose a novel FS-PCS model. While previous methods are based on feature optimization by mainly refining support features to enhance prototypes our method is based on correlation optimization referred to as Correlation Optimization Segmentation (COSeg). Specifically we compute Class-specific Multi-prototypical Correlation (CMC) for each query point representing its correlations to category prototypes. Then we propose the Hyper Correlation Augmentation (HCA) module to enhance CMC. Furthermore tackling the inherent property of few-shot training to incur base susceptibility for models we propose to learn non-parametric prototypes for the base classes during training. The learned base prototypes are used to calibrate correlations for the background class through a Base Prototypes Calibration (BPC) module. Experiments on popular datasets demonstrate the superiority of COSeg over existing methods. The code is available at github.com/ZhaochongAn/COSeg.
****
SEED-Bench: Benchmarking Multimodal Large Language Models
Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13299-13308
Multimodal large language models (MLLMs) building upon the foundation of powerful large language models (LLMs) have recently demonstrated exceptional capabilities in generating not only texts but also images given interleaved multimodal inputs (acting like a combination of GPT-4V and DALL-E 3). However existing MLLM benchmarks remain limited to assessing only models' comprehension ability of single image-text inputs failing to keep up with the strides made in MLLMs. A comprehensive benchmark is imperative for investigating the progress and uncovering the limitations of current MLLMs. In this work we categorize the capabilities of MLLMs into hierarchical levels from L_0 to L_4 based on the modalities they can accept and generate and propose SEED-Bench a comprehensive benchmark that evaluates the hierarchical capabilities of MLLMs. Specifically SEED-Bench comprises 24K multiple-choice questions with accurate human annotations which spans 27 dimensions including the evaluation of both text and image generation. Multiple-choice questions with groundtruth options derived from human annotation enables an objective and efficient assessment of model performance eliminating the need for human or GPT intervention during evaluation. We further evaluate the performance of 22 prominent open-source MLLMs and summarize valuable observations. By revealing the limitations of existing MLLMs through extensive evaluations we aim for SEED-Bench to provide insights that will motivate future research towards the goal of General Artificial Intelligence.
****
BrainWash: A Poisoning Attack to Forget in Continual Learning
Ali Abbasi, Parsa Nooralinejad, Hamed Pirsiavash, Soheil Kolouri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24057-24067
Continual learning has gained substantial attention within the deep learning community offering promising solutions to the challenging problem of sequential learning. Yet a largely unexplored facet of this paradigm is its susceptibility to adversarial attacks especially with the aim of inducing forgetting. In this paper we introduce "BrainWash" a novel data poisoning method tailored to impose forgetting on a continual learner. By adding the BrainWash noise to a variety of baselines we demonstrate how a trained continual learner can be induced to forget its previously learned tasks catastrophically even when using these continual learning baselines. An important feature of our approach is that the attacker requires no access to previous tasks' data and is armed merely with the model's current parameters and the data belonging to the most recent task. Our extensive experiments highlight the efficacy of BrainWash showcasing degradation in performance across various regularization and memory replay-based continual learning methods. Our code is available here: https://github.com/mint-vu/Brainwash
****
GreedyViG: Dynamic Axial Graph Construction for Efficient Vision GNNs
Mustafa Munir, William Avery, Md Mostafijur Rahman, Radu Marculescu; Proceedings

 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
024, pp. 6118-6127
Vision graph neural networks (ViG) offer a new avenue for exploration in compute
r vision. A major bottleneck in ViGs is the inefficient k-nearest neighbor (KNN)
 operation used for graph construction. To solve this issue we propose a new met
hod for designing ViGs Dynamic Axial Graph Construction (DAGC) which is more eff
icient than KNN as it limits the number of considered graph connections made wit
hin an image. Additionally we propose a novel CNN-GNN architecture GreedyViG whi
ch uses DAGC. Extensive experiments show that GreedyViG beats existing ViG CNN a
nd ViT architectures in terms of accuracy GMACs and parameters on image classifi
cation object detection instance segmentation and semantic segmentation tasks. O
ur smallest model GreedyViG-S achieves 81.1% top-1 accuracy on ImageNet-1K 2.9%
higher than Vision GNN and 2.2% higher than Vision HyperGraph Neural Network (Vi
HGNN) with less GMACs and a similar number of parameters. Our largest model Gree
dyViG-B obtains 83.9% top-1 accuracy 0.2% higher than Vision GNN with a 66.6% de
crease in parameters and a 69% decrease in GMACs. GreedyViG-B also obtains the s
ame accuracy as ViHGNN with a 67.3% decrease in parameters and a 71.3% decrease
in GMACs. Our work shows that hybrid CNN-GNN architectures not only provide a ne
w avenue for designing efficient models but that they can also exceed the perfor
mance of current state-of-the-art models.
****
Relightable and Animatable Neural Avatar from Sparse-View Video
Zhen Xu, Sida Peng, Chen Geng, Linzhan Mou, Zihan Yan, Jiaming Sun, Hujun Bao, X
iaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2024, pp. 990-1000
This paper tackles the problem of creating relightable and animatable neural ava
tars from sparse-view (or monocular) videos of dynamic humans under unknown illu
mination. Previous neural human reconstruction methods produce animatable avatar
s from sparse views using deformed Signed Distance Fields (SDF) but are non-reli
ghtable. While differentiable inverse rendering methods have succeeded in the ma
terial recovery of static objects it is not straightforward to extend them to dy
namic humans since it is computationally intensive to compute pixel-surface inte
rsection and light visibility on deformed SDFs for relighting. To solve this cha
llenge we propose a Hierarchical Distance Query (HDQ) algorithm to approximate t
he world space SDF under arbitrary human poses. Specifically we estimate coarse
SDF based on a parametric human model and compute fine SDF by exploiting the inv
ariance of SDF w.r.t. local deformation. Based on HDQ we leverage sphere tracing
 to efficiently estimate the surface intersection and light visibility. This all
ows us to develop the first system to recover relightable and animatable neural
avatars from sparse or monocular inputs. Experiments show that our approach prod
uces superior results compared to state-of-the-art methods. Our project page is
available at https://zju3dv.github.io/relightable_avatar.
****
FreePoint: Unsupervised Point Cloud Instance Segmentation
Zhikai Zhang, Jian Ding, Li Jiang, Dengxin Dai, Guisong Xia; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp.
 28254-28263
Instance segmentation of point clouds is a crucial task in 3D field with numerou
s applications that involve localizing and segmenting objects in a scene. Howeve
r achieving satisfactory results requires a large number of manual annotations w
hich is time-consuming and expensive. To alleviate dependency on annotations we
propose a novel framework FreePoint for underexplored unsupervised class-agnosti
c instance segmentation on point clouds. In detail we represent the point featur
es by combining coordinates colors and self-supervised deep features. Based on t
he point features we perform a bottom-up multicut algorithm to segment point clo
uds into coarse instance masks as pseudo labels which are used to train a point
cloud instance segmentation model. We propose an id-as-feature strategy at this
stage to alleviate the randomness of the multicut algorithm and improve the pseu
do labels' quality. During training we propose a weakly-supervised two-step trai
ning strategy and corresponding losses to overcome the inaccuracy of coarse mask

s. FreePoint has achieved breakthroughs in unsupervised class-agnostic instance segmentation on point clouds and outperformed previous traditional methods by over 18.2% and a competitive concurrent work UnScene3D by 5.5% in AP. Additionally when used as a pretext task and fine-tuned on S3DIS FreePoint performs significantly better than existing self-supervised pre-training methods with limited annotations and surpasses CSC by 6.0% in AP with 10% annotation masks. Code will be released at https://github.com/zzk273/FreePoint.
****
Pose Adapted Shape Learning for Large-Pose Face Reenactment
Gee-Sern Jison Hsu, Jie-Ying Zhang, Huang Yu Hsiang, Wei-Jie Hong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7413-7422
We propose the Pose Adapted Shape Learning (PASL) for large-pose face reenactment. The PASL framework consists of three modules namely the Pose-Adapted face Encoder (PAE) the Cycle-consistent Shape Generator (CSG) and the Attention-Embedded Generator (AEG). Different from previous approaches that use a single face encoder for identity preservation we propose multiple Pose-Adapted face Encodes (PAEs) to better preserve facial identity across large poses. Given a source face and a reference face the CSG generates a recomposed shape that fuses the source identity and reference action in the shape space and meets the cycle consistency requirement. Taking the shape code and the source as inputs the AEG learns the attention within the shape code and between the shape code and source style to enhance the generation of the desired target face. As existing benchmark datasets are inappropriate for evaluating large-pose face reenactment we propose a scheme to compose large-pose face pairs and introduce the MPIE-LP (Large Pose) and VoxCeleb2-LP datasets as the new large-pose benchmarks. We compared our approach with state-of-the-art methods on MPIE-LP and VoxCeleb2-LP for large-pose performance and on VoxCeleb1 for the common scope of pose variation.
****
Object Pose Estimation via the Aggregation of Diffusion Features
Tianfu Wang, Guosheng Hu, Hongguang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10238-10247
Estimating the pose of objects from images is a crucial task of 3D scene understanding and recent approaches have shown promising results on very large benchmarks. However these methods experience a significant performance drop when dealing with unseen objects. We believe that it results from the limited generalizability of image features. To address this problem we have an in-depth analysis on the features of diffusion models e.g. Stable Diffusion which hold substantial potential for modeling unseen objects. Based on this analysis we then innovatively introduce these diffusion features for object pose estimation. To achieve this we propose three distinct architectures that can effectively capture and aggregate diffusion features of different granularity greatly improving the generalizability of object pose estimation. Our approach outperforms the state-of-the-art methods by a considerable margin on three popular benchmark datasets LM O-LM and T-LESS. In particular our method achieves higher accuracy than the previous best arts on unseen objects: 98.2% vs. 93.5% on Unseen LM 85.9% vs. 76.3% on Unseen O-LM showing the strong generalizability of our method. Our code is released at https://github.com/Tianfu18/diff-feats-pose.
****
Circuit Design and Efficient Simulation of Quantum Inner Product and Empirical Studies of Its Effect on Near-Term Hybrid Quantum-Classic Machine Learning
Hao Xiong, Yehui Tang, Xinyu Ye, Junchi Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26162-26170
For the essential operation namely inner product (IP) as widely adopted in classic computing e.g. matrix multiplication its quantum counterpart: quantum inner product (QIP) has also been recently theoretically explored with a verifiable lower complexity on quantum computers. However it remains unclear for the embodiment of the quantum circuits (QC) for QIP let alone a (thorough) evaluation of the QIP circuits especially in a practical context in the NISQ era by applying QIP to ML via hybrid quantum-classic pipelines. In this paper we carefully design the

QIP circuits from scratch whose complexity is in accordance with the theoretica l complexity. To make the simulation tractable on classic computers especially w hen it is integrated in the gradient-based hybrid ML pipelines we further devise a highly-efficient simulation scheme by directly simulates the output state. Ex periments show that the scheme accelerates the simulation for more than 68k time s compared with the previous circuit simulator. This allows our empirical evalua tion on typical machine learning tasks ranging from supervised and self-supervis ed learning via neural nets to K-Means clustering. The results show that the cal culation error brought by typical quantum mechanisms would incur in general litt le influence on the final numerical results given sufficient qubits. However cer tain tasks e.g. ranking in K-Means could be more sensitive to quantum noise.
****
How to Make Cross Encoder a Good Teacher for Efficient Image-Text Retrieval?
Yuxin Chen, Zongyang Ma, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Ying Shan, Xiaojuan Qi, Weiming Hu; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26994-27003
Dominant dual-encoder models enable efficient image-text retrieval but suffer fr om limited accuracy while the cross-encoder models offer higher accuracy at the expense of efficiency. Distilling cross-modality matching knowledge from cross-e ncoder to dual-encoder provides a natural approach to harness their strengths. T hus we investigate the following valuable question: how to make cross-encoder a good teacher for dual-encoder? Our findings are threefold: (1) Cross-modal simil arity score distribution of cross-encoder is more concentrated while the result of dual-encoder is nearly normal making vanilla logit distillation less effectiv e. However ranking distillation remains practical as it is not affected by the s core distribution. (2) Only the relative order between hard negatives conveys va lid knowledge while the order information between easy negatives has little sign ificance. (3) Maintaining the coordination between distillation loss and dual-en coder training loss is beneficial for knowledge transfer. Based on these finding s we propose a novel Contrastive Partial Ranking Distillation (CPRD) method whic h implements the objective of mimicking relative order between hard negative sam ples with contrastive learning. This approach coordinates with the training of t he dual-encoder effectively transferring valid knowledge from the cross-encoder to the dual-encoder. Extensive experiments on image-text retrieval and ranking t asks show that our method surpasses other distillation methods and significantly improves the accuracy of dual-encoder.
****
Diffeomorphic Template Registration for Atmospheric Turbulence Mitigation
Dong Lao, Congli Wang, Alex Wong, Stefano Soatto; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25107-2511 6
We describe a method for recovering the irradiance underlying a collection of im ages corrupted by atmospheric turbulence. Since supervised data is often technic ally impossible to obtain assumptions and biases have to be imposed to solve thi s inverse problem and we choose to model them explicitly. Rather than initializi ng a latent irradiance ("template") by heuristics to estimate deformation we sel ect one of the images as a reference and model the deformation in this image by the aggregation of the optical flow from it to other images exploiting a prior i mposed by Central Limit Theorem. Then with a novel flow inversion module the mod el registers each image TO the template but WITHOUT the template avoiding artifa cts related to poor template initialization. To illustrate the robustness of the method we simply (i) select the first frame as the reference and (ii) use the s implest optical flow to estimate the warpings yet the improvement in registratio n is decisive in the final reconstruction as we achieve state-of-the-art perform ance despite its simplicity. The method establishes a strong baseline that can b e further improved by integrating it seamlessly into more sophisticated pipeline s or with domain-specific methods if so desired.
****
Selective Nonlinearities Removal from Digital Signals
Krzysztof A. Maliszewski, Magdalena A. Urba?ska, Varvara Vetrova, Sylwia M. Kole

nderska; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25028-25036
Many instruments performing optical and non-optical imaging and sensing such as Optical Coherence Tomography (OCT) Magnetic Resonance Imaging or Fourier-transform spectrometry produce digital signals containing modulations sine-like components which only after Fourier transformation give information about the structure or characteristics of the investigated object. Due to the fundamental physics-related limitations of such methods the distribution of these signal components is often nonlinear and when not properly compensated leads to the resolution precision or quality drop in the final image. Here we propose an innovative approach that has the potential to allow cleaning of the signal from the nonlinearities but most of all it now allows to switch the given order off leaving all others intact. The latter provides a tool for more in-depth analysis of the nonlinearity-inducing properties of the investigated object which can lead to applications in early disease detection or more sensitive sensing of chemical compounds. We consider OCT signals and nonlinearities up to the third order. In our approach we propose two neural networks: one to remove solely the second-order nonlinearity and the other for removing solely the third-order nonlinearity. The input of the networks is a novel two-dimensional data structure with all the information needed for the network to infer a nonlinearity-free signal. We describe the developed networks and present the results for second-order and third-order nonlinearity removal in OCT data representing the images of various objects: a mirror glass and fruits.
****
NB-GTR: Narrow-Band Guided Turbulence Removal
Yifei Xia, Chu Zhou, Chengxuan Zhu, Minggui Teng, Chao Xu, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24934-24943
The removal of atmospheric turbulence is crucial for long-distance imaging. Leveraging the stochastic nature of atmospheric turbulence numerous algorithms have been developed that employ multi-frame input to mitigate the turbulence. However when limited to a single frame existing algorithms face substantial performance drops particularly in diverse real-world scenes. In this paper we propose a robust solution to turbulence removal from an RGB image under the guidance of an additional narrow-band image broadening the applicability of turbulence mitigation techniques in real-world imaging scenarios. Our approach exhibits a substantial suppression in the magnitude of turbulence artifacts by using only a pair of images thereby enhancing the clarity and fidelity of the captured scene.
****
Can Biases in ImageNet Models Explain Generalization?
Paul Gavrikov, Janis Keuper; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22184-22194
The robust generalization of models to rare in-distribution (ID) samples drawn from the long tail of the training distribution and to out-of-training-distribution (OOD) samples is one of the major challenges of current deep learning methods. For image classification this manifests in the existence of adversarial attacks the performance drops on distorted images and a lack of generalization to concepts such as sketches. The current understanding of generalization in neural networks is very limited but some biases that differentiate models from human vision have been identified and might be causing these limitations. Consequently several attempts with varying success have been made to reduce these biases during training to improve generalization. We take a step back and sanity-check these attempts. Fixing the architecture to the well-established ResNet-50 we perform a large-scale study on 48 ImageNet models obtained via different training methods to understand how and if these biases - including shape bias spectral biases and critical bands - interact with generalization. Our extensive study results reveal that contrary to previous findings these biases are insufficient to accurately predict the generalization of a model holistically. We provide access to all checkpoints and evaluation code at https://github.com/paulgavrikov/biases_vs_generalization/

****
NRDF: Neural Riemannian Distance Fields for Learning Articulated Pose Priors
Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, Gerard Pons-Moll; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2024, pp. 1661-1671
Faithfully modeling the space of articulations is a crucial task that allows rec
overy and generation of realistic poses and remains a notorious challenge. To th
is end we introduce Neural Riemannian Distance Fields (NRDFs) data-driven priors
 modeling the space of plausible articulations represented as the zero-level-set
 of a neural field in a high-dimensional product-quaternion space. To train NRDF
s only on positive examples we introduce a new sampling algorithm ensuring that
the geodesic distances follow a desired distribution yielding a principled dista
nce field learning paradigm. We then devise a projection algorithm to map any ra
ndom pose onto the level-set by an adaptive-step Riemannian optimizer adhering t
o the product manifold of joint rotations at all times. NRDFs can compute the Ri
emannian gradient via backpropagation and by mathematical analogy are related to
 Riemannian flow matching a recent generative model. We conduct a comprehensive
evaluation of NRDF against other pose priors in various downstream tasks i.e. po
se generation image-based pose estimation and solving inverse kinematics highlig
hting NRDF's superior performance. Besides humans NRDF's versatility extends to
hand and animal poses as it can effectively represent any articulation.
****
RepAn: Enhanced Annealing through Re-parameterization
Xiang Fei, Xiawu Zheng, Yan Wang, Fei Chao, Chenglin Wu, Liujuan Cao; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2024, pp. 5798-5808
The simulated annealing algorithm aims to improve model convergence through mult
iple restarts of training. However existing annealing algorithms overlook the co
rrelation between different cycles neglecting the potential for incremental lear
ning. We contend that a fixed network structure prevents the model from recogniz
ing distinct features at different training stages. To this end we propose RepAn
 redesigning the irreversible re-parameterization (Rep) method and integrating i
t with annealing to enhance training. Specifically the network goes through Rep
expansion restoration and backpropagation operations during training and iterati
ng through these processes in each annealing round. Such a method exhibits good
generalization and is easy to apply and we provide theoretical explanations for
its effectiveness. Experiments demonstrate that our method improves baseline per
formance by 6.38% on the CIFAR-100 dataset and 2.80% on ImageNet achieving state
-of-the-art performance in the Rep field. The code is available at https://githu
b.com/xfey/RepAn.
****
Generative Quanta Color Imaging
Vishal Purohit, Junjie Luo, Yiheng Chi, Qi Guo, Stanley H. Chan, Qiang Qiu; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2024, pp. 25138-25148
The astonishing development of single-photon cameras has created an unprecedente
d opportunity for scientific and industrial imaging. However the high data throu
ghput generated by these 1-bit sensors creates a significant bottleneck for low-
power applications. In this paper we explore the possibility of generating a col
or image from a single binary frame of a single-photon camera. We evidently find
 this problem being particularly difficult to standard colorization approaches d
ue to the substantial degree of exposure variation. The core innovation of our p
aper is an exposure synthesis model framed under a neural ordinary differential
equation (Neural ODE) that allows us to generate a continuum of exposures from a
 single observation. This innovation ensures consistent exposure in binary image
s that colorizers take on resulting in notably enhanced colorization. We demonst
rate applications of the method in single-image and burst colorization and show
superior generative performance over baselines. Project website can be found at
https://vishal-s-p.github.io/projects/2023/generative_quanta_color.html
****

Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers
Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, Sergey Tulyakov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13320-13331
The quality of the data and annotation upper-bounds the quality of a downstream model. While there exist large text corpora and image-text pairs high-quality video-text data is much harder to collect. First of all manual labeling is more time-consuming as it requires an annotator to watch an entire video. Second videos have a temporal dimension consist of a number of scenes stacked together and show multiple actions. Accordingly to establish a video dataset with high-quality captions we propose an automatic approach leveraging multimodal inputs such as textual video description subtitles and individual video frames. Specifically we curate 3.8M high-resolution videos from the publicly available HD-VILA-100M dataset. We then split them into semantically consistent video clips and apply multiple cross-modality teacher models to obtain captions for each video. Next we finetune a retrieval model on a small subset where the best caption of each video is manually selected and then employ the model in the whole dataset to select the best caption as the annotation. In this way we get 70M videos paired with high-quality text captions. We dub the dataset as Panda-70M. We show the value of the proposed dataset on three downstream tasks: video captioning video and text retrieval and text-driven video generation. The models trained on the proposed data score substantially better on the majority of metrics across all the tasks.
****
Overload: Latency Attacks on Object Detection for Edge Devices
Erh-Chung Chen, Pin-Yu Chen, I-Hsin Chung, Che-Rung Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 24716-24725
Nowadays the deployment of deep learning-based applications is an essential task owing to the increasing demands on intelligent services. In this paper we investigate latency attacks on deep learning applications. Unlike common adversarial attacks for misclassification the goal of latency attacks is to increase the inference time which may stop applications from responding to the requests within a reasonable time. This kind of attack is ubiquitous for various applications and we use object detection to demonstrate how such kind of attacks work. We also design a framework named Overload to generate latency attacks at scale. Our method is based on a newly formulated optimization problem and a novel technique called spatial attention. This attack serves to escalate the required computing costs during the inference time consequently leading to an extended inference time for object detection. It presents a significant threat especially to systems with limited computing resources. We conducted experiments using YOLOv5 models on Nvidia NX. Compared to existing methods our method is simpler and more effective. The experimental results show that with latency attacks the inference time of a single image can be increased ten times longer in reference to the normal setting. Moreover our findings pose a potential new threat to all object detection tasks requiring non-maximum suppression (NMS) as our attack is NMS-agnostic.
****
DreamControl: Control-Based Text-to-3D Generation with 3D Self-Prior
Tianyu Huang, Yihan Zeng, Zhilu Zhang, Wan Xu, Hang Xu, Songcen Xu, Rynson W.H. Lau, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5364-5373
3D generation has raised great attention in recent years. With the success of text-to-image diffusion models the 2D-lifting technique becomes a promising route to controllable 3D generation. However these methods tend to present inconsistent geometry which is also known as the Janus problem. We observe that the problem is caused mainly by two aspects i.e. viewpoint bias in 2D diffusion models and overfitting of the optimization objective. To address it we propose a two-stage 2D-lifting framework namely DreamControl which optimizes coarse NeRF scenes as 3D self-prior and then generates fine-grained objects with control-based score distillation. Specifically adaptive viewpoint sampling and boundary integrity metr

ic are proposed to ensure the consistency of generated priors. The priors are th
en regarded as input conditions to maintain reasonable geometries in which condi
tional LoRA and weighted score are further proposed to optimize detailed texture
s. DreamControl can generate high-quality 3D content in terms of both geometry c
onsistency and texture fidelity. Moreover our control-based optimization guidanc
e is applicable to more downstream tasks including user-guided generation and 3D
 animation.
****
Infrared Small Target Detection with Scale and Location Sensitivity
Qiankun Liu, Rui Liu, Bolun Zheng, Hongkui Wang, Ying Fu; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17
490-17499
Recently infrared small target detection (IRSTD) has been dominated by deep-lear
ning-based methods. However these methods mainly focus on the design of complex
model structures to extract discriminative features leaving the loss functions f
or IRSTD under-explored. For example the widely used Intersection over Union (Io
U) and Dice losses lack sensitivity to the scales and locations of targets limit
ing the detection performance of detectors. In this paper we focus on boosting d
etection performance with a more effective loss but a simpler model structure. S
pecifically we first propose a novel Scale and Location Sensitive (SLS) loss to
handle the limitations of existing losses: 1) for scale sensitivity we compute a
 weight for the IoU loss based on target scales to help the detector distinguish
 targets with different scales: 2) for location sensitivity we introduce a penal
ty term based on the center points of targets to help the detector localize targ
ets more precisely. Then we design a simple Multi-Scale Head to the plain U-Net
(MSHNet). By applying SLS loss to each scale of the predictions our MSHNet outpe
rforms existing state-of-the-art methods by a large margin. In addition the dete
ction performance of existing detectors can be further improved when trained wit
h our SLS loss demonstrating the effectiveness and generalization of our SLS los
s. The code is available at https://github.com/ying-fu/MSHNet.
****
Self-supervised Debiasing Using Low Rank Regularization
Geon Yeong Park, Chanyong Jung, Sangmin Lee, Jong Chul Ye, Sang Wan Lee; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2024, pp. 12395-12405
Spurious correlations can cause strong biases in deep neural networks impairing
generalization ability. While most existing debiasing methods require full super
vision on either spurious attributes or target labels training a debiased model
from a limited amount of both annotations is still an open question. To address
this issue we investigate an interesting phenomenon using the spectral analysis
of latent representations: spuriously correlated attributes make neural networks
 inductively biased towards encoding lower effective rank representations. We al
so show that a rank regularization can amplify this bias in a way that encourage
s highly correlated features. Leveraging these findings we propose a self-superv
ised debiasing framework potentially compatible with unlabeled samples. Specific
ally we first pretrain a biased encoder in a self-supervised manner with the ran
k regularization serving as a semantic bottleneck to enforce the encoder to lear
n the spuriously correlated attributes. This biased encoder is then used to disc
over and upweight bias-conflicting samples in a downstream task serving as a boo
sting to effectively debias the main model. Remarkably the proposed debiasing fr
amework significantly improves the generalization performance of self-supervised
 learning baselines and in some cases even outperforms state-of-the-art supervis
ed debiasing approaches.
****
ODIN: A Single Model for 2D and 3D Segmentation
Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W. Harley, Gabriel Sarch,
Kriti Aggarwal, Vishrav Chaudhary, Katerina Fragkiadaki; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 356
4-3574
State-of-the-art models on contemporary 3D segmentation benchmarks like ScanNet

consume and label dataset-provided 3D point clouds obtained through post processing of sensed multiview RGB-D images. They are typically trained in-domain forego large-scale 2D pre-training and outperform alternatives that featurize the posed RGB-D multiview images instead. The gap in performance between methods that consume posed images versus post-processed 3D point clouds has fueled the belief that 2D and 3D perception require distinct model architectures. In this paper we challenge this view and propose ODIN (Omni-Dimensional INstance segmentation) a model that can segment and label both 2D RGB images and 3D point clouds using a transformer architecture that alternates between 2D within-view and 3D cross-view information fusion. Our model differentiates 2D and 3D feature operations through the positional encodings of the tokens involved which capture pixel coordinates for 2D patch tokens and 3D coordinates for 3D feature tokens. ODIN achieves state-of-the-art performance on ScanNet200 Matterport3D and AI2THOR 3D instance segmentation benchmarks and competitive performance on ScanNet S3DIS and COCO. It outperforms all previous works by a wide margin when the sensed 3D point cloud is used in place of the point cloud sampled from 3D mesh. When used as the 3D perception engine in an instructable embodied agent architecture it sets a new state-of-the-art on the TEACh action-from-dialogue benchmark. Our code and checkpoints can be found at the project website: https://odin-seg.github.io.
****
SD4Match: Learning to Prompt Stable Diffusion Model for Semantic Matching
Xinghui Li, Jingyi Lu, Kai Han, Victor Adrian Prisacariu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27558-27568
In this paper we address the challenge of matching semantically similar keypoints across image pairs. Existing research indicates that the intermediate output of the UNet within the Stable Diffusion (SD) can serve as robust image feature maps for such a matching task. We demonstrate that by employing a basic prompt tuning technique the inherent potential of Stable Diffusion can be harnessed resulting in a significant enhancement in accuracy over previous approaches. We further introduce a novel conditional prompting module that conditions the prompt on the local details of the input image pairs leading to a further improvement in performance. We designate our approach as SD4Match short for Stable Diffusion for Semantic Matching. Comprehensive evaluations of SD4Match on the PF-Pascal PF-Willow and SPair-71k datasets show that it sets new benchmarks in accuracy across all these datasets. Particularly SD4Match outperforms the previous state-of-the-art by a margin of 12 percentage points on the challenging SPair-71k dataset. Code is available at the project website: https://sd4match.active.vision.
****
InitNO: Boosting Text-to-Image Diffusion Models via Initial Noise Optimization
Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, Di Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9380-9389
Recent strides in the development of diffusion models exemplified by advancements such as Stable Diffusion have underscored their remarkable prowess in generating visually compelling images. However the imperative of achieving a seamless alignment between the generated image and the provided prompt persists as a formidable challenge. This paper traces the root of these difficulties to invalid initial noise and proposes a solution in the form of Initial Noise Optimization (InitNO) a paradigm that refines this noise. Considering text prompts not all random noises are effective in synthesizing semantically-faithful images. We design the cross-attention response score and the self-attention conflict score to evaluate the initial noise bifurcating the initial latent space into valid and invalid sectors. A strategically crafted noise optimization pipeline is developed to guide the initial noise towards valid regions. Our method validated through rigorous experimentation shows a commendable proficiency in generating images in strict accordance with text prompts. Our code is available at https://github.com/xiefan-guo/initno.
****
Neural Video Compression with Feature Modulation

Jiahao Li, Bin Li, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2024, pp. 26099-26108
The emerging conditional coding-based neural video codec (NVC) shows superiority
 over commonly-used residual coding-based codec and the latest NVC already claim
s to outperform the best traditional codec. However there still exist critical p
roblems blocking the practicality of NVC. In this paper we propose a powerful co
nditional coding-based NVC that solves two critical problems via feature modulat
ion. The first is how to support a wide quality range in a single model. Previou
s NVC with this capability only supports about 3.8 dB PSNR range on average. To
tackle this limitation we modulate the latent feature of the current frame via t
he learnable quantization scaler. During the training we specially design the un
iform quantization parameter sampling mechanism to improve the harmonization of
encoding and quantization. This results in a better learning of the quantization
 scaler and helps our NVC support about 11.4 dB PSNR range. The second is how to
 make NVC still work under a long prediction chain. We expose that the previous
SOTA NVC has an obvious quality degradation problem when using a large intra-per
iod setting. To this end we propose modulating the temporal feature with a perio
dically refreshing mechanism to boost the quality. Notably under single intra-fr
ame setting our codec can achieve 29.7% bitrate saving over previous SOTA NVC wi
th 16% MACs reduction. Our codec serves as a notable landmark in the journey of
NVC evolution. The codes are at https://github.com/microsoft/DCVC.
****
Data Poisoning based Backdoor Attacks to Contrastive Learning
Jinghuai Zhang, Hongbin Liu, Jinyuan Jia, Neil Zhenqiang Gong; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p
p. 24357-24366
Contrastive learning (CL) pre-trains general-purpose encoders using an unlabeled
 pre-training dataset which consists of images or image-text pairs. CL is vulner
able to data poisoning based backdoor attacks (DPBAs) in which an attacker injec
ts poisoned inputs into the pre-training dataset so the encoder is backdoored. H
owever existing DPBAs achieve limited effectiveness. In this work we take the fi
rst step to analyze the limitations of existing backdoor attacks and propose new
 DPBAs called CorruptEncoder to CL. CorruptEncoder introduces a new attack strat
egy to create poisoned inputs and uses a theory-guided method to maximize attack
 effectiveness. Our experiments show that CorruptEncoder substantially outperfor
ms existing DPBAs. In particular CorruptEncoder is the first DPBA that achieves
more than 90% attack success rates with only a few (3) reference images and a sm
all poisoning ratio (0.5%). Moreover we also propose a defense called localized
cropping to defend against DPBAs. Our results show that our defense can reduce t
he effectiveness of DPBAs but it sacrifices the utility of the encoder highlight
ing the need for new defenses.
****
Multimodal Sense-Informed Forecasting of 3D Human Motions
Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, Hong Zhou; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2
144-2154
Predicting future human pose is a fundamental application for machine intelligen
ce which drives robots to plan their behavior and paths ahead of time to seamles
sly accomplish human-robot collaboration in real-world 3D scenarios. Despite enc
ouraging results existing approaches rarely consider the effects of the external
 scene on the motion sequence leading to pronounced artifacts and physical impla
usibilities in the predictions. To address this limitation this work introduces
a novel multi-modal sense-informed motion prediction approach which conditions h
igh-fidelity generation on two modal information: external 3D scene and internal
 human gaze and is able to recognize their salience for future human activity. F
urthermore the gaze information is regarded as the human intention and combined
with both motion and scene features we construct a ternary intention-aware atten
tion to supervise the generation to match where the human wants to reach. Meanwh
ile we introduce semantic coherence-aware attention to explicitly distinguish th
e salient point clouds and the underlying ones to ensure a reasonable interactio

n of the generated sequence with the 3D scene. On two real-world benchmarks the proposed method achieves state-of-the-art performance both in 3D human pose and trajectory prediction. More detailed results are available on the page: https://sites.google.com/view/cvpr2024sif3d.
****
FlowerFormer: Empowering Neural Architecture Encoding using a Flow-aware Graph Transformer
Dongyeong Hwang, Hyunju Kim, Sunwoo Kim, Kijung Shin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6128-6137
The success of a specific neural network architecture is closely tied to the dataset and task it tackles; there is no one-size-fits-all solution. Thus considerable efforts have been made to quickly and accurately estimate the performances of neural architectures without full training or evaluation for given tasks and datasets. Neural architecture encoding has played a crucial role in the estimation and graphbased methods which treat an architecture as a graph have shown prominent performance. For enhanced representation learning of neural architectures we introduce FlowerFormer a powerful graph transformer that incorporates the information flows within a neural architecture. FlowerFormer consists of two key components: (a) bidirectional asynchronous message passing inspired by the flows; (b) global attention built on flow-based masking. Our extensive experiments demonstrate the superiority of FlowerFormer over existing neural encoding methods and its effectiveness extends beyond computer vision models to include graph neural networks and auto speech recognition models. Our code is available at http://github.com/y0ngjaenius/CVPR2024_FLOWERFormer.
****
EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models
Jingyuan Yang, Jiawei Feng, Hui Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6358-6368
Recent years have witnessed remarkable progress in image generation task where users can create visually astonishing images with high-quality. However exsiting text-to-image diffusion models are proficient in generating concrete concepts (dogs) but encounter challenges with more abstract ones (emotions). Several efforts have been made to modify image emotions with color and style adjustments facing limitations in effectively conveying emotions with fixed image contents. In this work we introduce Emotional Image Content Generation (EIGC) a new task to generate semantic-clear and emotion-faithful images given emotion categories. Specifically we propose an emotion space and construct a mapping network to align it with powerful Contrastive Language-Image Pre-training (CLIP) space providing a concrete interpretation of abstract emotions. Attribute loss and emotion confidence are further proposed to ensure the semantic diversity and emotion fidelity of the generated images. Our method outperforms the state-the-art text-to-image approaches both quantitatively and qualitatively where we derive three custom metrics i.e.emotion accuracy semantic clarity and semantic diversity. In addition to generation our method can help emotion understanding and inspire emotional art design. Project page: https://vcc.tech/research/2024/EmoGen.
****
Finding Lottery Tickets in Vision Models via Data-driven Spectral Foresight Pruning
Leonardo Iurada, Marco Ciccone, Tatiana Tommasi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16142-16151
Recent advances in neural network pruning have shown how it is possible to reduce the computational costs and memory demands of deep learning models before training. We focus on this framework and propose a new pruning at initialization algorithm that leverages the Neural Tangent Kernel (NTK) theory to align the training dynamics of the sparse network with that of the dense one. Specifically we show how the usually neglected data-dependent component in the NTK's spectrum can be taken into account by providing an analytical upper bound to the NTK's trace obtained by decomposing neural networks into individual paths. This leads to our Path eXclusion (PX) a foresight pruning method designed to preserve the paramet

ers that mostly influence the NTK's trace. PX is able to find lottery tickets (i
.e. good paths) even at high sparsity levels and largely reduces the need for ad
ditional training. When applied to pre-trained models it extracts subnetworks di
rectly usable for several downstream tasks resulting in performance comparable t
o those of the dense counterpart but with substantial cost and computational sav
ings.
****
InNeRF360: Text-Guided 3D-Consistent Object Inpainting on 360-degree Neural Radi
ance Fields
Dongqing Wang, Tong Zhang, Alaa Abboud, Sabine Süsstrunk; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12
677-12686
We propose InNeRF360 an automatic system that accurately removes text-specified
objects from 360-degree Neural Radiance Fields (NeRF). The challenge is to effec
tively remove objects while inpainting perceptually consistent content for the m
issing regions which is particularly demanding for existing NeRF models due to t
heir implicit volumetric representation. Moreover unbounded scenes are more pron
e to floater artifacts in the inpainted region than frontal-facing scenes as the
 change of object appearance and background across views is more sensitive to in
accurate segmentations and inconsistent inpainting. With a trained NeRF and a te
xt description our method efficiently removes specified objects and inpaints vis
ually consistent content without artifacts. We apply depth-space warping to enfo
rce consistency across multiview text-encoded segmentations and then refine the
inpainted NeRF model using perceptual priors and 3D diffusion-based geometric pr
iors to ensure visual plausibility. Through extensive experiments in segmentatio
n and inpainting on 360-degree and frontal-facing NeRFs we show that InNeRF360 i
s effective and enhances NeRF's editability. Project page: https://ivrl.github.i
o/InNeRF360.
****
Neural Implicit Representation for Building Digital Twins of Unknown Articulated
 Objects
Yijia Weng, Bowen Wen, Jonathan Tremblay, Valts Blukis, Dieter Fox, Leonidas Gui
bas, Stan Birchfield; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2024, pp. 3141-3150
We address the problem of building digital twins of unknown articulated objects
from two RGBD scans of the object at different articulation states. We decompose
 the problem into two stages each addressing distinct aspects. Our method first
reconstructs object-level shape at each state then recovers the underlying artic
ulation model including part segmentation and joint articulations that associate
 the two states. By explicitly modeling point-level correspondences and exploiti
ng cues from images 3D reconstructions and kinematics our method yields more acc
urate and stable results compared to prior work. It also handles more than one m
ovable part and does not rely on any object shape or structure priors. Project p
age: https://github.com/NVlabs/DigitalTwinArt
****
Progressive Semantic-Guided Vision Transformer for Zero-Shot Learning
Shiming Chen, Wenjin Hou, Salman Khan, Fahad Shahbaz Khan; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2
3964-23974
Zero-shot learning (ZSL) recognizes the unseen classes by conducting visual-sema
ntic interactions to transfer semantic knowledge from seen classes to unseen one
s supported by semantic information (e.g. attributes). However existing ZSL meth
ods simply extract visual features using a pre-trained network backbone (i.e. CN
N or ViT) which fail to learn matched visual-semantic correspondences for repres
enting semantic-related visual features as lacking of the guidance of semantic i
nformation resulting in undesirable visual-semantic interactions. To tackle this
 issue we propose a progressive semantic-guided vision transformer for zero-shot
 learning (dubbed ZSLViT). ZSLViT mainly considers two properties in the whole n
etwork: i) discover the semantic-related visual representations explicitly and i
i) discard the semantic-unrelated visual information. Specifically we first intr

oduce semantic-embedded token learning to improve the visual-semantic correspond
ences via semantic enhancement and discover the semantic-related visual tokens e
xplicitly with semantic-guided token attention. Then we fuse low semantic-visual
 correspondence visual tokens to discard the semantic-unrelated visual informati
on for visual enhancement. These two operations are integrated into various enco
ders to progressively learn semantic-related visual representations for accurate
 visual-semantic interactions in ZSL. The extensive experiments show that our ZS
LViT achieves significant performance gains on three popular benchmark datasets
i.e. CUB SUN and AWA2.
****
IS-Fusion: Instance-Scene Collaborative Fusion for Multimodal 3D Object Detectio
n
Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, We
nguan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2024, pp. 14905-14915
Bird's eye view (BEV) representation has emerged as a dominant solution for desc
ribing 3D space in autonomous driving scenarios. However objects in the BEV repr
esentation typically exhibit small sizes and the associated point cloud context
is inherently sparse which leads to great challenges for reliable 3D perception.
 In this paper we propose IS-Fusion an innovative multimodal fusion framework th
at jointly captures the Instance- and Scene-level contextual information. IS-Fus
ion essentially differs from existing approaches that only focus on the BEV scen
e-level fusion by explicitly incorporating instance-level multimodal information
 thus facilitating the instance-centric tasks like 3D object detection. It compr
ises a Hierarchical Scene Fusion (HSF) module and an Instance-Guided Fusion (IGF
) module. HSF applies Point-to-Grid and Grid-to-Region transformers to capture t
he multimodal scene context at different granularities. IGF mines instance candi
dates explores their relationships and aggregates the local multimodal context f
or each instance. These instances then serve as guidance to enhance the scene fe
ature and yield an instance-aware BEV representation. On the challenging nuScene
s benchmark IS-Fusion outperforms all the published multimodal works to date.
****
Building Bridges across Spatial and Temporal Resolutions: Reference-Based Super-
Resolution via Change Priors and Conditional Diffusion Model
Runmin Dong, Shuai Yuan, Bin Luo, Mengxuan Chen, Jinxiao Zhang, Lixian Zhang, We
ijia Li, Juepeng Zheng, Haohuan Fu; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2024, pp. 27684-27694
Reference-based super-resolution (RefSR) has the potential to build bridges acro
ss spatial and temporal resolutions of remote sensing images. However existing R
efSR methods are limited by the faithfulness of content reconstruction and the e
ffectiveness of texture transfer in large scaling factors. Conditional diffusion
 models have opened up new opportunities for generating realistic high-resolutio
n images but effectively utilizing reference images within these models remains
an area for further exploration. Furthermore content fidelity is difficult to gu
arantee in areas without relevant reference information. To solve these issues w
e propose a change-aware diffusion model named Ref-Diff for RefSR using the land
 cover change priors to guide the denoising process explicitly. Specifically we
inject the priors into the denoising model to improve the utilization of referen
ce information in unchanged areas and regulate the reconstruction of semanticall
y relevant content in changed areas. With this powerful guidance we decouple the
 semantics-guided denoising and reference texture-guided denoising processes to
improve the model performance. Extensive experiments demonstrate the superior ef
fectiveness and robustness of the proposed method compared with state-of-the-art
 RefSR methods in both quantitative and qualitative evaluations. The code and da
ta are available at https://github.com/dongrunmin/RefDiff.
****
Vanishing-Point-Guided Video Semantic Segmentation of Driving Scenes
Diandian Guo, Deng-Ping Fan, Tongyu Lu, Christos Sakaridis, Luc Van Gool; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2024, pp. 3544-3553

The estimation of implicit cross-frame correspondences and the high computational cost have long been major challenges in video semantic segmentation (VSS) for driving scenes. Prior works utilize keyframes feature propagation or cross-frame attention to address these issues. By contrast we are the first to harness vanishing point (VP) priors for more effective segmentation. Intuitively objects near VPs (i.e. away from the vehicle) are less discernible. Moreover they tend to move radially away from the VP over time in the usual case of a forward-facing camera a straight road and linear forward motion of the vehicle. Our novel efficient network for VSS named VPSeg incorporates two modules that utilize exactly this pair of static and dynamic VP priors: sparse-to-dense feature mining (DenseVP) and VP-guided motion fusion (MotionVP). MotionVP employs VP-guided motion estimation to establish explicit correspondences across frames and help attend to the most relevant features from neighboring frames while DenseVP enhances weak dynamic features in distant regions around VPs. These modules operate within a context-detail framework which separates contextual features from high-resolution local features at different input resolutions to reduce computational costs. Contextual and local features are integrated through contextualized motion attention (CMA) for the final prediction. Extensive experiments on two popular driving segmentation benchmarks Cityscapes and ACDC demonstrate that VPSeg outperforms previous SOTA methods with only modest computational overhead.
****
Enhancing Intrinsic Features for Debiasing via Investigating Class-Discerning Common Attributes in Bias-Contrastive Pair
Jeonghoon Park, Chaeyeon Chung, Jaegul Choo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12332-12341
In the image classification task deep neural networks frequently rely on bias attributes that are spuriously correlated with a target class in the presence of dataset bias resulting in degraded performance when applied to data without bias attributes. The task of debiasing aims to compel classifiers to learn intrinsic attributes that inherently define a target class rather than focusing on bias attributes. While recent approaches mainly focus on emphasizing the learning of data samples without bias attributes (i.e. bias-conflicting samples) compared to samples with bias attributes (i.e. bias-aligned samples) they fall short of directly guiding models where to focus for learning intrinsic features. To address this limitation this paper proposes a method that provides the model with explicit spatial guidance that indicates the region of intrinsic features. We first identify the intrinsic features by investigating the class-discerning common features between a bias-aligned (BA) sample and a bias-conflicting (BC) sample (i.e. bias-contrastive pair). Next we enhance the intrinsic features in the BA sample that are relatively under-exploited for prediction compared to the BC sample. To construct the bias-contrastive pair without using bias information we introduce a bias-negative score that distinguishes BC samples from BA samples employing a biased model. The experiments demonstrate that our method achieves state-of-the-art performance on synthetic and real-world datasets with various levels of bias severity.
****
LAMP: Learn A Motion Pattern for Few-Shot Video Generation
Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7089-7098
In this paper we present a few-shot text-to-video framework LAMP which enables a text-to-image diffusion model to Learn A specific Motion Pattern with 8 16 videos on a single GPU. Unlike existing methods which require a large number of training resources or learn motions that are precisely aligned with template videos it achieves a trade-off between the degree of generation freedom and the resource costs for model training. Specifically we design a motion-content decoupled pipeline that uses an off-the-shelf text-to-image model for content generation so that our tuned video diffusion model mainly focuses on motion learning. The well-developed text-to-image techniques can provide visually pleasing and diverse content as generation conditions which highly improves video quality and generatio

n freedom. To capture the features of temporal dimension we expand the pre-trained 2D convolution layers of the T2I model to our novel temporal-spatial motion learning layers and modify the attention blocks to the temporal level. Additionally we develop an effective inference trick shared-noise sampling which can improve the stability of videos without computational costs. Our method can also be flexibly applied to other tasks e.g. real-world image animation and video editing. Extensive experiments demonstrate that LAMP can effectively learn the motion pattern on limited data and generate high-quality videos. The code and models are available at https://rq-wu.github.io/projects/LAMP.
****
Compositional Chain-of-Thought Prompting for Large Multimodal Models
Chancharik Mitra, Brandon Huang, Trevor Darrell, Roei Herzig; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14420-14431
The combination of strong visual backbones and Large Language Model (LLM) reasoning has led to Large Multimodal Models (LMMs) becoming the current standard for a wide range of vision and language (VL) tasks. However recent research has shown that even the most advanced LMMs still struggle to capture aspects of compositional visual reasoning such as attributes and relationships between objects. One solution is to utilize scene graphs (SGs)---a formalization of objects and their relations and attributes that has been extensively used as a bridge between the visual and textual domains. Yet scene graph data requires scene graph annotations which are expensive to collect and thus not easily scalable. Moreover finetuning an LMM based on SG data can lead to catastrophic forgetting of the pretraining objective. To overcome this inspired by chain-of-thought methods we propose Compositional Chain-of-Thought (CCoT) a novel zero-shot Chain-of-Thought prompting method that utilizes SG representations in order to extract compositional knowledge from an LMM. Specifically we first generate an SG using the LMM and then use that SG in the prompt to produce a response. Through extensive experiments we find that the proposed CCoT approach not only improves LMM performance on several vision and language VL compositional benchmarks but also improves the performance of several popular LMMs on general multimodal benchmarks without the need for fine-tuning or annotated ground-truth SGs. Code: https://github.com/chancharikmitra/CCoT
****
Diffusion Time-step Curriculum for One Image to 3D Generation
Xuanyu Yi, Zike Wu, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9948-9958
Score distillation sampling (SDS) has been widely adopted to overcome the absence of unseen views in reconstructing 3D objects from a single image. It leverages pre-trained 2D diffusion models as teacher to guide the reconstruction of student 3D models. Despite their remarkable success SDS-based methods often encounter geometric artifacts and texture saturation. We find out the crux is the overlooked indiscriminate treatment of diffusion time-steps during optimization: it unreasonably treats the student-teacher knowledge distillation to be equal at all time-steps and thus entangles coarse-grained and fine-grained modeling. Therefore we propose the Diffusion Time-step Curriculum one-image-to-3D pipeline (DTC123) which involves both the teacher and student models collaborating with the time-step curriculum in a coarse-to-fine manner. Extensive experiments on NeRF4 RealFusion15 GSO and Level50 benchmark demonstrate that DTC123 can produce multi-view consistent high-quality and diverse 3D assets. Codes and more generation demos will be released in https://github.com/yxymessi/DTC123.
****
Language-driven Object Fusion into Neural Radiance Fields with Pose-Conditioned Dataset Updates
Ka Chun Shum, Jaeyeon Kim, Binh-Son Hua, Duc Thanh Nguyen, Sai-Kit Yeung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5176-5187
Neural radiance field (NeRF) is an emerging technique for 3D scene reconstructio

n and modeling. However current NeRF-based methods are limited in the capabilities of adding or removing objects. This paper fills the aforementioned gap by proposing a new language-driven method for object manipulation in NeRFs through dataset updates. Specifically to insert an object represented by a set of multi-view images into a background NeRF we use a text-to-image diffusion model to blend the object into the given background across views. The generated images are then used to update the NeRF so that we can render view-consistent images of the object within the background. To ensure view consistency we propose a dataset update strategy that prioritizes the radiance field training based on camera poses in a pose-ordered manner. We validate our method in two case studies: object insertion and object removal. Experimental results show that our method can generate photo-realistic results and achieves state-of-the-art performance in NeRF editing.
****
Adaptive Hyper-graph Aggregation for Modality-Agnostic Federated Learning
Fan Qi, Shuai Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12312-12321
In Federated Learning (FL) the issue of statistical data heterogeneity has been a significant challenge to the field's ongoing development. This problem is further exacerbated when clients' data vary in modalities. In response to these issues of statistical heterogeneity and modality incompatibility we propose the Adaptive Hyper-graph Aggregation framework a novel solution for Modality-Agnostic Federated Learning. We design a Modular Architecture for Local Model with single modality setting the stage for efficient intra-modality sharing and inter-modality complementarity. An innovative Global Consensus Prototype Enhancer is crafted to assimilate and broadcast global consensus knowledge within the network. At the core of our approach lies the Adaptive Hyper-graph Learning Strategy which effectively tackles the inherent challenges of modality incompatibility and statistical heterogeneity within federated learning environments accomplishing this adaptively even without the server being aware of the clients' modalities. Our approach tested on three multimodal benchmark datasets demonstrated strong performance across diverse data distributions affirming its effectiveness in multimodal federated learning.
****
SPIN: Simultaneous Perception Interaction and Navigation
Shagun Uppal, Ananye Agarwal, Haoyu Xiong, Kenneth Shaw, Deepak Pathak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18133-18142
While there has been remarkable progress recently in the fields of manipulation and locomotion mobile manipulation remains a long-standing challenge. Compared to locomotion or static manipulation a mobile system must make a diverse range of long-horizon tasks feasible in unstructured and dynamic environments. While the applications are broad and interesting there are a plethora of challenges in developing these systems such as coordination between the base and arm reliance on onboard perception for perceiving and interacting with the environment and most importantly simultaneously integrating all these parts together. Prior works approach the problem using disentangled modular skills for mobility and manipulation that are trivially tied together. This causes several limitations such as compounding errors delays in decision-making and no whole-body coordination. In this work we present a reactive mobile manipulation framework that uses an active visual system to consciously perceive and react to its environment. Similar to how humans leverage whole-body and hand-eye coordination we develop a mobile manipulator that exploits its ability to move and see more specifically -- to move in order to see and to see in order to move. This allows it to not only move around and interact with its environment but also choose "when" to perceive "what" using an active visual system. We observe that such an agent learns to navigate around complex cluttered scenarios while displaying agile whole-body coordination using only ego-vision without needing to create environment maps. Videos are available at https://spin-robot.github.io
****

DREAM: Diffusion Rectification and Estimation-Adaptive Models
Jinxin Zhou, Tianyu Ding, Tianyi Chen, Jiachen Jiang, Ilya Zharkov, Zhihui Zhu,
Luming Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2024, pp. 8342-8351
We present DREAM a novel training framework representing Diffusion Rectification
 and Estimation-Adaptive Models requiring minimal code changes (just three lines
) yet significantly enhancing the alignment of training with sampling in diffusi
on models. DREAM features two components: diffusion rectification which adjusts
training to reflect the sampling process and estimation adaptation which balance
s perception against distortion. When applied to image super-resolution (SR) DRE
AM adeptly navigates the tradeoff between minimizing distortion and preserving h
igh image quality. Experiments demonstrate DREAM's superiority over standard dif
fusion-based SR methods showing a to faster training convergence and a to reduct
ion in necessary sampling steps to achieve comparable or superior results. We ho
pe DREAM will inspire a rethinking of diffusion model training paradigms.
****
Exploring the Potential of Large Foundation Models for Open-Vocabulary HOI Detec
tion
Ting Lei, Shaofeng Yin, Yang Liu; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2024, pp. 16657-16667
Open-vocabulary human-object interaction (HOI) detection which is concerned with
 the problem of detecting novel HOIs guided by natural language is crucial for u
nderstanding human-centric scenes. However prior zero-shot HOI detectors often e
mploy the same levels of feature maps to model HOIs with varying distances leadi
ng to suboptimal performance in scenes containing human-object pairs with a wide
 range of distances. In addition these detectors primarily rely on category name
s and overlook the rich contextual information that language can provide which i
s essential for capturing open vocabulary concepts that are typically rare and n
ot well-represented by category names alone. In this paper we introduce a novel
end-to-end open vocabulary HOI detection framework with conditional multi-level
decoding and fine-grained semantic enhancement (CMD-SE) harnessing the potential
 of Visual-Language Models (VLMs). Specifically we propose to model human-object
 pairs with different distances with different levels of feature maps by incorpo
rating a soft constraint during the bipartite matching process. Furthermore by l
everaging large language models (LLMs) such as GPT models we exploit their exten
sive world knowledge to generate descriptions of human body part states for vari
ous interactions. Then we integrate the generalizable and fine-grained semantics
 of human body parts to improve interaction recognition. Experimental results on
 two datasets SWIG-HOI and HICO-DET demonstrate that our proposed method achieve
s state-of-the-art results in open vocabulary HOI detection. The code and models
 are available at https://github.com/ltttpku/CMD-SE-release.
****